DYNAMICS OF PROTEIN-DRUG INTERACTIONS INFERRED FROM STRUCTURAL ENSEMBLES AND PHYSICS-BASED MODELS

by

Ahmet Bakan

BS, Chemistry, Koç University, 2005

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Ahmet Bakan

It was defended on

December 03, 2009

and approved by

Dr. John S. Lazo, Professor, Department of Pharmacology

Dr. Xiang-Qun Xie, Professor, Department of Pharmaceutical Sciences

Dr. Chakra Chennubhotla, Assistant Professor, Department of Computational Biology

Dr. Christopher J. Langmead, Assistant Professor, Department of Computer Science, CMU

Dissertation Advisor: Dr. Ivet Bahar, Professor, Department of Computational Biology

DYNAMICS OF PROTEIN-DRUG INTERACTIONS INFERRED FROM STRUCTURAL ENSEMBLES AND PHYSICS-BASED MODELS

Ahmet Bakan, PhD

University of Pittsburgh, 2009

The conformational flexibility of target proteins is a major challenge in understanding and modeling protein-drug interactions. A fundamental issue, yet to be clarified, is whether the observed conformational changes are controlled by the protein, or induced by the inhibitor. While the concept of induced fit has been widely adopted for describing the structural changes that accompany ligand binding, there is growing evidence in support of the dominance of proteins' intrinsic dynamics, which has been evolutionarily optimized to accommodate its functional interactions. The wealth of structural data for target proteins in the presence of different ligands now permits us to make a critical assessment of the balance between these two effects in selecting the bound forms. We focused on three widely studied drug targets, HIV-1 reverse transcriptase, p38 MAP kinase, and cyclin-dependent kinase 2. A total of 292 structures determined for these enzymes in the presence of different inhibitors as well as unbound form permitted us to perform an extensive comparative analysis of the conformational space accessed upon ligand binding, and its relation to the intrinsic dynamics prior to ligand binding as predicted by elastic network model analysis. Further, we analyzed NMR ensembles of ubiquitin and calmodulin representing their microseconds range solution dynamics. Our results show that the ligand selects the conformer that best matches its structural and dynamic properties amongst the

conformers intrinsically accessible to the protein in the unliganded form. The results suggest that simple but robust rules encoded in the protein structure play a dominant role in pre-defining the mechanisms of ligand binding, which may be advantageously exploited in designing inhibitors. We apply these lessons to the study of MAP kinase phosphatases (MKPs), which are therapeutically relevant but challenging signaling enzymes. Our study provides insights into the interactions and selectivity of MKP inhibitors and shows how an allosteric inhibition mechanism holds for a recently discovered inhibitor of MKP-3. We also provide evidence for the functional significance of the structure-encoded dynamics of rhodopsin and nicotinic acetylcholine receptor, members of two membrane proteins classes serving as targets for more than 40% of all current FDA approved drugs.

TABLE OF CONTENTS

PREFACE ..... **XIV**

1.0 INTRODUCTION **1**

1.1 ROLE OF PROTEIN DYNAMICS IN SMALL MOLECULE RECOGNITION AND BINDING **1**

1.2 COMPUTATIONAL METHODS FOR LEARNING AND MODELING PROTEIN DYNAMICS AND INTERACTIONS **4**

1.3 DUAL-SPECIFICITY MITOGEN-ACTIVATED PROTEIN (MAP) KINASE PHOSPHATASES (MKPS) **7**

1.4 MEMBRANE PROTEINS **11**

1.5 SPECIFIC AIMS AND OUTLINE OF THE DISSERTATION **13**

2.0 THEORY AND METHODS **17**

2.1 PRINCIPAL COMPONENT ANALYSIS (PCA) OF STRUCTURAL ENSEMBLES **18**

2.1.1 Iterative optimal superimposition of structures **19**

2.1.2 Covariance matrix: a measure of correlations between residue fluctuations **21**

2.1.3 Projection of conformations onto the subspace spanned by the PCs **23**

2.1.4 A note on the analysis of X-ray ensembles **24**

2.2 A NORMAL MODE ANALYSIS BASED ON ANISOTROPIC NETWORK MODEL (ANM) 25

2.2.1 Normal Mode Analysis (NMA) 26

2.2.1.1 Hessian matrix 27

2.2.1.2 Normal modes 29

2.2.1.3 Interpretation of Normal Modes: Significance of slow modes 30

2.2.1.4 Covariance computed from NMA: Bridging with PCA of structural ensembles 32

2.2.2 ANM: Theory and Foundations 33

2.2.2.1 Assumptions and Model Parameters 33

2.2.2.2 Potential Function and the Hessian 34

2.2.2.3 Generation of Alternative Conformations Using the ANM 36

2.3 COMPARISON OF ESSENTIAL MODES DERIVED FROM PCA AND ANM ANALYSIS 37

2.4 ALL-ATOM MODELS OF PROTEINS 39

2.4.1 Force Fields 40

2.4.2 Molecular modeling and simulation packages 44

2.4.3 Comparative modeling 45

2.5 ALTERNATIVE PROTEIN CONFORMATIONS FROM NORMAL MODE GUIDED DISPLACEMENTS 47

2.6 MOLECULAR DOCKING 49

2.6.1 Unbiased docking simulations using AutoDock 50

2.6.2 Focused docking simulations using GOLD 52

2.6.3 The need for generating large ensembles of docking poses and post-docking clustering analysis 54
2.7 SUPPLEMENTARY METHODS 56
2.7.1 Electrostatic potential calculations using APBS 56
2.7.2 Scientific programming using Python 57
3.0 THE INTRINSIC DYNAMICS OF ENZYMES PLAYS A DOMINANT ROLE IN DETERMINING THE STRUCTURAL CHANGES OBSERVED IN INHIBITOR BOUND STRUCTURES 60
3.1 ANALYSIS OF X-RAY STRUCTURAL ENSEMBLES FOR ENZYMES TARGETED BY DRUGS 60
3.1.1 Structural Ensembles of drug target enzymes 62
3.1.2 HIV-1 reverse transcriptase 65
3.1.3 p38 MAP kinase 72
3.1.4 Cyclin-dependent kinase 2 77
3.2 SOLUTION DYNAMICS OF PROTEINS COMPARED WITH ANM 81
3.3 DISCUSSION 89
4.0 TARGETING MAP KINASE PHOSPHATASES (MKPS): INSIGTS FROM STRUCTURE-BASED MODELING OF INHIBITOR INTERACTIONS 93
4.1 INTRODUCTION 93
4.1.1 Sequence, Structure, and Function 93
4.1.2 Catalytic activation of MKPs upon substrate recognition offers alternative inhibition mechanisms 95
4.2 MKP-1/INHIBITOR INTERACTIONS 97

4.2.1 Structurally unique inhibitors of MKP-1 from a focus library of pyrrole carboxamides 97
4.2.2 Basis of specificity of inhibitors from electrostatic surface potential calculations 98
4.2.3 Interactions of MKP-1 with pyrrole carboxamides 100
4.3 A NOVEL MKP-3 INHIBITOR FROM ZEBRAFISH CHEMICAL SCREENS 103
4.3.1 Zebrafish chemical screens identify a novel MKP-3 inhibitor 103
4.3.2 Putative binding site and inhibition mechanism from modeling 104
4.3.2.1 Identification of potential bindings sites using unbiased docking 105
4.3.2.2 Flexible docking for a detailed assessment of potential inhibition mechanisms 107
4.3.3 Experimental testing of the hypothesis 112
4.4 DISCUSSION 114
5.0 UNDERSTANDING FUNCTIONAL MECHANISMS OF MEMBRANE PROTEINS 117
5.1 INTRODUCTION 117
5.2 RHODOPSIN 119
5.2.1 G-protein couple receptors 119
5.2.2 Rhodopsin 120
5.2.3 PCA of rhodopsin structure ensemble 122
5.2.4 Correspondence between ANM modes and PCA modes 124
5.3 NICOTINIC ACETYLCHOLINE RECEPTOR 127

5.3.1 Ligand-gated ion channels 127
5.3.2 Nicotinic acetylcholine receptor (nAChR) structure 127
5.3.3 Models of nAChR channel gating 129
5.3.4 Recent structures confirm quaternary twist-to-open mode 132
6.0 CONCLUSION AND FUTURE WORK 134
6.1 PROTEIN RECOGNITION DYNAMICS AND SMALL-MOLECULE BINDING 134
6.1.1 Why aren't some ANM modes observed in experimental datasets? 134
6.1.2 Potential improvements to ANM-guided conformer generation 136
6.2 MKP INHIBITORS 139
6.2.1 An iterative approach for designing more potent BCI analogs 139
6.2.2 Characterization of the putative BCI binding site 141
6.2.3 Estimating the druggability of the putative BCI binding site 141
6.2.4 Virtual screening for new chemotypes 142
APPENDIX A 144
BIBLIOGRAPHY 146

LIST OF TABLES

Table 3.1 Overlap between PCA modes obtained from complete ensembles of reverse transcriptase (RT) structures with those obtained from NNRTI bound subset of the ensemble . 70

Table 3.2 Overlap between PCA and ANM modes for the ensemble of RT structures 70

Table 3.3 Overlap between PCA and ANM modes for the ensemble of p38 structures 74

Table 3.4 Overlap between PCA and ANM modes for the ensemble of Cdk2 structures..... 80

Table 3.5 Overlap between PCA and ANM modes for NMR ensembles 87

Table 3.6 Cumulative Overlap..... 90

Table 4.1 MKP catalytic domain structures and their sequence identities. 94

Table A.1 PDB IDs of RT structures. 144

Table A.2 PDB IDs of p38 structures. 145

Table A.3 PDB IDs of Cdk2 structures. 145

LIST OF FIGURES

Figure 1.1 Range of equilibrium motions that accompany small-molecule ligand binding. 2

Figure 1.2 Energy profile of the native state modeled at different resolutions..... 6

Figure 1.3 Comparison of the active sites of MKP-5, PTP1B and Cdk2. 8

Figure 1.4 Distribution of biological targets among approved drugs. 9

Figure 1.5 Distribution of small molecule drugs based on the targeted molecular function. 12

Figure 2.1 Schematic representation of a protein configuration. 20

Figure 3.1 RMSD distributions for three sets of structures: HIV-1 RT, p38 and Cdk2. 63

Figure 3.2 Distribution of pair-wise Tanimoto coefficients for inhibitors in X-ray datasets. 64

Figure 3.3 PCA and ANM results for HIV-1 RT..... 66

Figure 3.4 Projection of conformational changes onto PC1. 68

Figure 3.5 Top three ANM modes for reference RT structure. 71

Figure 3.6 PCA and ANM results for p38 MAP kinase. 73

Figure 3.7 Top three ANM modes for reference p38 structure. 77

Figure 3.8 PCA and ANM results for Cdk2. 78

Figure 3.9 Top three ANM modes for reference Cdk2 structure..... 81

Figure 3.10 RMSD distributions for three ensembles of NMR models. 83

Figure 3.11 Comparison of the PC and ANM modes for ubiquitin and CaM ensembles. 85

Figure 3.12 Comparison of the PC and ANM modes for CaM ensembles..... 86

Figure 3.13 Directions of PCA modes and corresponding ANM modes. 88

Figure 3.14 Schematic description of observed mechanism. 91

Figure 4.1 MKP CD domain structures in active and inactive states. 94

Figure 4.2 Alignment of MKP catalytic domain sequences. 96

Figure 4.3 MKP-1 inhibitor from a focused pyrrole carboxamide library. 98

Figure 4.4 Surface properties of dual-specificity phosphatases (DSPs) and potential inhibitor binding sites. 99

Figure 4.5 CD structures and docking solutions for MKP-1 inhibitors. 101

Figure 4.6 BCI structure (A) and zebrafish embryos before (B) and after (C) BCI treatment. .. 103

Figure 4.7 Results from unbiased docking simulations of BCI. 105

Figure 4.8 MKP-3 general acid loop (encircled region) makes crystal contacts. 108

Figure 4.9 Combined ANM mode direction (A) and alternative MKP-3 conformations along this mode (B). 109

Figure 4.10 BCI interactions predicted using focused and flexible docking. 111

Figure 4.11 BCI interactions with (A) MKP-1 and (B) VH3. 112

Figure 4.12 BCI does not inhibit OMFP phosphorylation. 113

Figure 4.13 BCI is an allosteric inhibitor of MKP-3. 113

Figure 4.14 Multiplicity of BCI binding modes. 115

Figure 5.1 Rhodopsin (A) (chromophore in magenta) and G-protein bound opsin* (B). 121

Figure 5.2 PCA of rhodopsin structural ensemble. 124

Figure 5.3 Opsin structures deformed along PC and ANM modes. 126

Figure 5.4 Ligand-gated ion channel nAChR structure and dynamics. 130

Figure 5.5 Comparison of open and closed bacterial ligand-gated ion channels (LGICs). 133

PREFACE

I gratefully thank my advisor Dr. Ivet Bahar for her support throughout my studies, and helping me in many ways to get accustomed to living in Pittsburgh. Her enthusiasm for exploring a broad range of biological systems has been inspirational to me. I feel privileged and fortunate to have worked under her guidance on several interdisciplinary research problems.

I am indebted to Dr. John S. Lazo for our pleasant and fruitful collaboration. I sincerely thank our collaborators Drs. Andreas Vogt, Michael Tsang, Billy W. Day, Kay M. Brummond, and Peter Wipf. It has been a very valuable and a unique experience for me to work with them.

I thank Dr. Chakra Chennubhotla for our stimulating discussions and for his constructive criticism. I am thankful to my dissertation committee members Drs. Xiang-Qun Xie and Christopher J. Langmead, for their insightful questions, comments, and suggestions that helped me shape my dissertation.

Finally, I thank my beloved wife Merve Kovan-Bakan for being a part of my life, and to my family for their unconditional love and support.

1.0 INTRODUCTION

1.1 ROLE OF PROTEIN DYNAMICS IN SMALL MOLECULE RECOGNITION AND BINDING

The dynamic nature of proteins plays a critical role in molecular recognition. Understanding the determinants of ligand-recognition and -binding dynamics is a major challenge with impact on drug discovery. Yet, progress in this field has been impeded by the complexity and specificity of interactions, the multiplicity of conformations accessible under equilibrium conditions, as well as insufficient data on the structure and energetics of protein-ligand interactions.

Figure 1.1 provides an overview of the broad range of equilibrium motions that are observed to accompany small-molecule ligand binding. These motions range from rotations of side-chains interacting with the ligand to globular motions engaging domains or subunits. The atomic scale details of such changes are explored using X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy techniques.

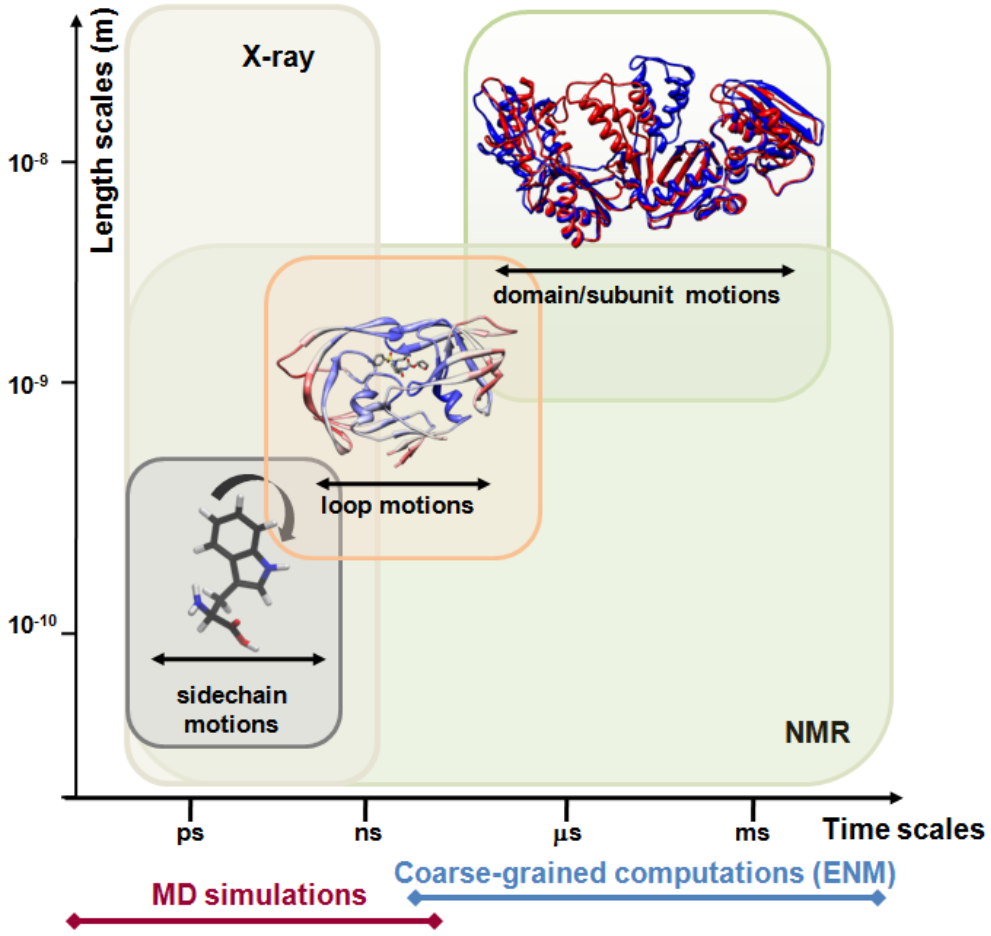


Figure 1.1 Range of equilibrium motions that accompany small-molecule ligand binding.

Motions that accompany the binding of small molecules range from side-chain rotations to slower concerted domain motions. X-ray crystallography and NMR are the primary sources of information on such conformational changes at atomic resolution. Also indicated along the abscissa are the timescales of processes that can be explored by molecular dynamics simulations (MD) and coarse-grained (CG) computations. Figure is adapted from (Bahar et al., 2009). Molecular diagrams here and in the following figures have been generated using Chimera (**http://www.cgl.ucsf.edu/chimera/**) (Pettersen et al., 2004), VMD (Humphrey et al., 1996), or PyMol (**http://pymol.org/**) visualization software.

Two different models have been proposed for explaining the conformational changes observed between the bound and unbound forms of proteins. The classical view, which dates back to the original work of Koshland in 1958, proposes an *induced fit* mechanism whereby

ligand binding induces conformational changes on the target protein (Koshland, 1958). Such an onset of conformational changes could be plausible on a *local* scale, i.e., slight rearrangements in side chain reorientations or even transitions between isomeric states could be triggered by the ligand. However, the more cooperative changes observed in other complexes, including concerted rearrangements of entire domains, have challenged this classical concept. The second, alternate view, pioneered by Monod, Wyman, and Changeux (MWC model) (Monod et al., 1965), has gained broad acceptance in the last decade, supported by experimental and computational studies (Ma et al., 1999; Kern & Zuiderweg, 2003; James et al., 2003; Tobi & Bahar, 2005; Mittermaier & Kay, 2006; Bahar et al., 2007; Showalter & Bruschweiler, 2007; Tang et al., 2007; Lange et al., 2008; Gsponer et al., 2008; Ivetac & McCammon, 2009). It is consistent with the accessibility of a host of conformational substates under native state conditions. Accordingly, the protein samples an ensemble of conformations (*pre-existing equilibrium*), a fraction of which is predisposed to recognize and bind a particular ligand (*conformational selection*). Therefore, observed structural rearrangements would not occur if it were not for the pre-disposition or *intrinsic dynamics* of the protein to fluctuate between multiple conformers including those prone to readily bind the ligand (Bahar et al., 2007).

A number of more recent studies suggest a more complex interplay between intrinsic dynamics and ligand-induced motions. For example, Okazaki and Takada reported that stronger and long-range interactions favor induced fit, while shorter-range interactions favor conformational selection (Okazaki & Takada, 2008). Even if binding occurs via conformational selection, additional rearrangements may be induced to stabilize the complex (James & Tawfik, 2005; Tobi & Bahar, 2005). And while protein-protein interactions may be strongly affected by

their intrinsic dynamics, it is unclear which effect, intrinsic dynamics *vs*. induced dynamics, plays a dominant role, in protein-small molecule interactions, which may entail in many cases highly specific, localized interactions. Sullivan and Holyoak argued for example that the presence of a lid at the binding site implies an induced fit mechanism (Sullivan & Holyoak, 2008); and folding upon binding is commonly observed in intrinsically disordered protein segments (Turjanski et al., 2008).

1.2 COMPUTATIONAL METHODS FOR LEARNING AND MODELING PROTEIN DYNAMICS AND INTERACTIONS

With the rapid accumulation of multiple liganded structures for a given protein in the Protein Data Bank (PDB) (Berman et al., 2000) and with the development of analytical models for rapid estimation of intrinsic dynamics, we are now in a position to (i) critically examine sets of conformations assumed by the same protein in the presence of different ligands and (ii) compare these conformational changes to those predicted for the unliganded protein using simplified, physics-based models. While such comparisons between experimental and computational data may be obscured by the heterogeneities of the accessible conformations and uncertainties in atomic coordinates, there exist powerful methods to extract dominant patterns from complex data.

With regard to experimental data, principal component analysis (PCA) is an old but powerful method to unveil the *principal* variations in structure. An excellent application is the recent examination of the ensemble of ubiquitin X-ray structures complexed with different substrates, in comparison to the ensemble of NMR models determined by residual dipolar coupling measurements (Lange et al., 2008). This study showed that the conformational changes assumed in different complexes, and those observed for the isolated protein in solution show close overlap, and essentially represent displacements along a well-defined (combined) principal mode of deformation intrinsically favored by the unbound protein.

As to structural dynamics, again a classical approach to retrieve dominant modes of motion is normal mode analysis (NMA) (Kitao & Go, 1999; Cui & Bahar, 2006). Normal mode analysis provides information on the equilibrium modes accessible to a system, assuming that the system is stabilized by harmonic potentials. Its application to proteins dates back to early 1980s. In recent years, NMA has seen a revival with the realization that highly simplified models, such as the anisotropic network model (ANM) (Atilgan et al., 2001; Eyal et al., 2006), can be utilized to efficiently predict global modes of motions (**Figure 1.2**). These motions are characterized by a high degree of collectivity, and usually lie at the lowest frequency end of the mode spectrum. They are insensitive to structural details or underlying force field, but defined by the overall architecture, or topology of inter-residue contacts in the native structure (Cui & Bahar, 2006; Nicolay & Sanejouand, 2006; Tama & Brooks, 2006).

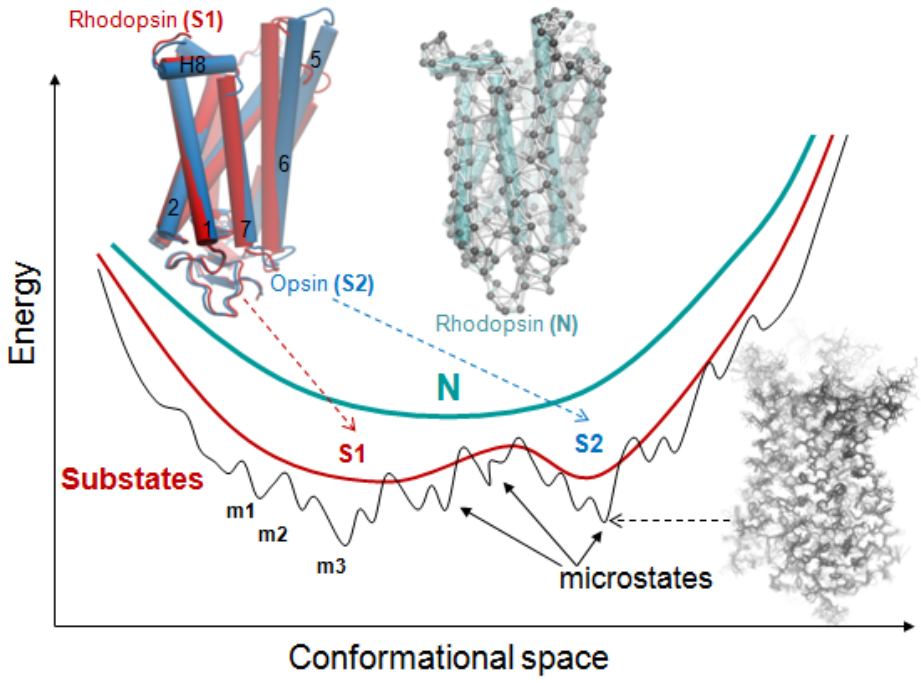


Figure 1.2 Energy profile of the native state modeled at different resolutions.

N denotes the native state, modeled at a CG scale as a single energy minimum. A detailed examination of the structure and energetics may reveal multiple substates (S1, S2, etc.), which in turn contain multiple microstates (m1, m2, etc.). Structural models corresponding to different hierarchical levels of resolution are shown: an elastic network model representation where the global energy minimum on a CG scale (N) is approximated by a harmonic potential along each mode direction (e.g. anisotropic network model (ANM)); two substates S1 and S2 sampled by global motions near native state conditions; and an ensemble of conformers sampled by small fluctuations in the neighborhood of each substate. The diagrams have been constructed using the following rhodopsin structures deposited in the PDB: 1U19 (N); 1U19 and 3CAP (S1 and S2) and 1F88, 1GZM, 1HZX, 1L9H, 1U19, 2G87, 2HPY, 2I35, 2I36, 2I37, 2J4Y, 2PED, 3C9L, 3C9M (microstates). Figure is adopted from (Bahar et al., 2009).

1.3 DUAL-SPECIFICITY MITOGEN-ACTIVATED PROTEIN (MAP) KINASE PHOSPHATASES (MKPS)

Mitogen-activated protein (MAP) kinase phosphatases (MKPs) are dual-specificity protein phosphatases (DSPs). They are important signal transduction enzymes that regulate various cellular processes, such as growth and differentiation, in opposition with protein kinases. MKPs form a family with 11 members under the protein tyrosine phosphatase (PTP) superfamily, which is distinguished by the active-site signature motif HCX_5R at the phosphatase (PTPase) loop (Alonso et al., 2004; Tonks, 2006).

MKPs dephosphorylate and inactivate MAP kinases (Theodosiou & Ashworth, 2002). Their substrates are p38 kinases (p38s), c-Jun amino-terminal kinases (JNKs) and extracellular signaling-related kinases (ERKs) (Jeffrey et al., 2007). Members of the MKP family show tissue specific expression patterns and tight growth/mitogen/stress regulated transcriptional induction profiles (Camps et al., 2000). At least five of these family members have been implicated with various cancer types (Wu, 2007). Among these, MKP-1, the prototypical member, and the structurally well-characterized MKP-3 have been most closely examined. MKP-1 is a nuclear protein and shows selectivity for p38s and JNKs over ERKs. Its expression levels are elevated in at least eight cancer types and correlate with the progression of breast (Wang et al., 2003) and lung (Vicent et al., 2004) cancers. MKP-3 is a cytosolic phosphatase with preference for ERKs over other MAP kinases. Overexpression of MKP-3 is associated with lung (Chen et al., 2007) and pancreatic (Furukawa et al., 2003) cancers. Yet undisclosed therapeutic potentials of MKPs have led to a growing cancer medicinal chemistry interest in exploring their biology.

MKPs, or in general DSPs, are defined by their ability to catalyze the removal of two covalently attached phosphate groups from tyrosine and serine/threonine residues on the same substrate. This activity is achieved via the PTPase catalytic mechanism at a relatively shallow active site (**Figure 1.3**); presumably, this shape is required for accommodating the phosphorylated serine/threonine residues which, in contrast to phosphorylated tyrosine, only extend slightly beyond the peptide backbone (Denu & Dixon, 1998). **Figure 1.3** illustrates the catalytic domain of MKP-5 in comparison to two extensively studied cell signaling enzymes, protein tyrosine phosphatase 1B (PTP1B) (Puius et al., 1997) and cyclin-dependent kinase 2 (Cdk2) (Bramson et al., 2001). PTP1B and Cdk2 structures have a deep groove at their active sites, which ensures the tight binding of small molecule ligands, and allows for the design of selective inhibitors. The lack of analogous active site features in MKPs increases the challenge for the design of small molecule inhibitors.

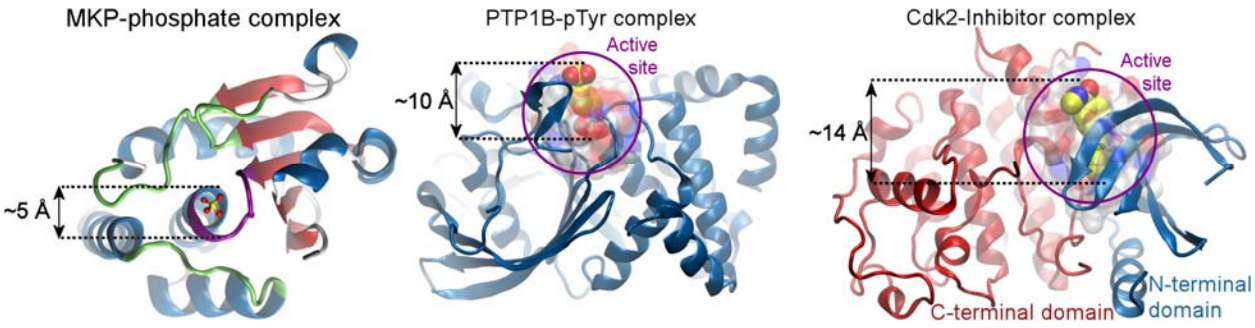


Figure 1.3 Comparison of the active sites of MKP-5, PTP1B and Cdk2.

MKP active sites are rather shallow, compared to the deep pockets that permit the insertion of ligands in PTP1B and Cdk2. Figure is adapted from (Bakan et al., 2008).

MKP inhibitors have been identified in natural product collections (Vogt et al., 2005), diversity-oriented chemical libraries (Vogt et al., 2003) as well as in larger scale drug-like compound libraries (Johnston et al., 2007). These compounds have not been widely used as molecular probes or lead compounds in part because of their lack of potency, redox properties, and inadequate phosphatase selectivity, despite some limited synthetic analog follow up studies (Lazo et al., 2006). **Figure 1.4** presents an overview of the ensemble of proteins targeted by approved drugs, retrieved from DrugBank (Wishart et al., 2008), which showcases the small fraction of protein phosphatases and the lack of representation of MKPs.

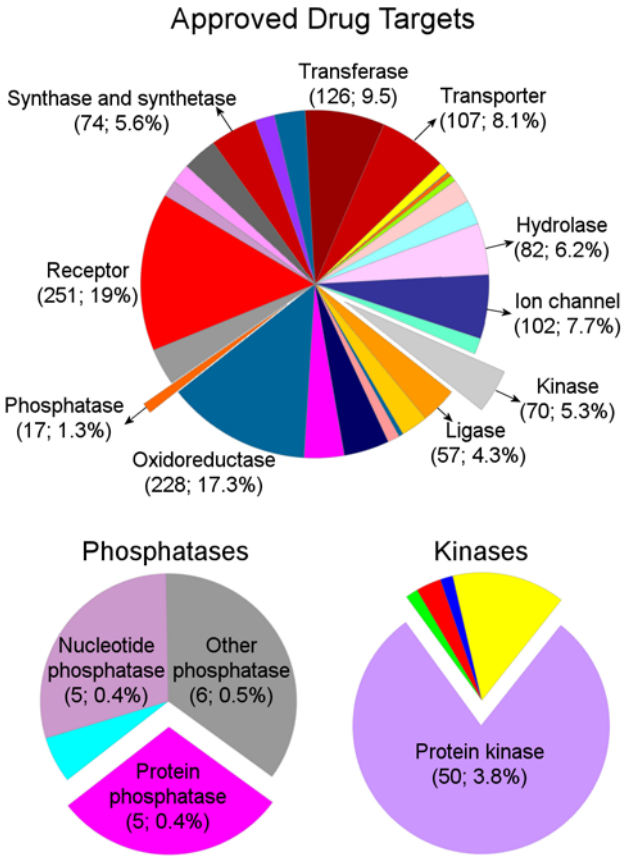


Figure 1.4 Distribution of biological targets among approved drugs.

Molecular functions of 1321 approved human drug targets retrieved from DrugBank (Wishart et al., 2008) are assigned using the PANTHER classification system (Thomas et al., 2003). Phosphatase and kinase slices are

enlarged in two separate pie charts. We note that a much smaller number of protein phosphatases are validated as targets compared to protein kinases, which performs the complementary chemical function. The numbers in parentheses represent the numbers of target proteins in each category and their fractional contribution to the entire set of approved drug targets. Protein tyrosine phosphatase (PTP) 1b, low molecular weight PTP, ser/thr-protein phosphatase 2A, tyrosine-protein phosphatase non-receptor types 4 are among protein phosphatases. There are no MKPs approved as drug targets. MKP-3 is listed in the DrugBank as an experimental drug target. Figure is adopted from (Bakan et al., 2008).

Computational methods are now ubiquitous in all aspects of drug discovery (Jorgensen, 2004). Particularly, structure-based modeling tools are broadly used in lead discovery and optimization (Shoichet et al., 2002). In the past three years, the number of known distinct MKP structures has more than doubled, providing the framework for developing structure-based modeling for compounds of therapeutic interest. It is reasonable to hypothesize that improving our understanding of the interactions of MKPs with their substrates and inhibitors could accelerate the discovery and development of therapeutic agents. Here we will focus on two therapeutically important MKPs, MKP-1 and MKP-3, and present results on their interactions with ligands at the molecular level.

We face two major challenges in the *in silico* design of lead compounds that target MKPs, which are common to most molecular docking efforts: (i) modeling the conformational flexibility of the protein and (ii) accounting for the entropic effects that stabilize the bound inhibitor conformations. While tackling these challenges, we apply the lessons that we learned from the analysis of protein-inhibitor complex structure ensembles. After presenting the results for MKPs, we finally discuss prospects toward addressing these challenges in general.

1.4 MEMBRANE PROTEINS

With their locations at cell boundaries, membrane proteins are involved not only in the transport of endogenous substrates/ions, but also in drug uptake (Dobson & Kell, 2008) and drug action. While approximately 30% of sequenced genes encode membrane proteins, the fraction of membrane proteins among drug targets has been estimated to be 70% in 2001 (Stahlberg et al., 2001). An updated distribution of drug targets is presented in **Figure 1.5.** The pie chart refers to 965 US Food and Drug Administration (FDA) approved small-molecule drugs, obtained from DrugBank (http://www.drugbank.ca) as primary source (Wishart et al., 2008), and refined using Therapeutic Target Database (DB) (Chen et al., 2002), SuperTarget DB (Gunther et al., 2008), and literature (Imming et al., 2006). A total of 380 proteins are targeted by these drugs, most of which belong to the human genome. The corresponding molecular functions, retrieved from the PANTHER Classification System (Thomas et al., 2003), are grouped into 71 functional categories. **Figure 1.5** displays the most frequently targeted ten such categories. The top-ranking four categories are G-protein coupled receptors (GPCRs), nuclear hormone receptors, ligand-gated ion channels (LGIC) and voltage-gated ion channels. These constitute targets for more than half of the drugs. These results are consistent with those recently compiled by Hopkins and coworkers (Overington et al., 2006), apart from minor differences presumably due to differences in the dataset.

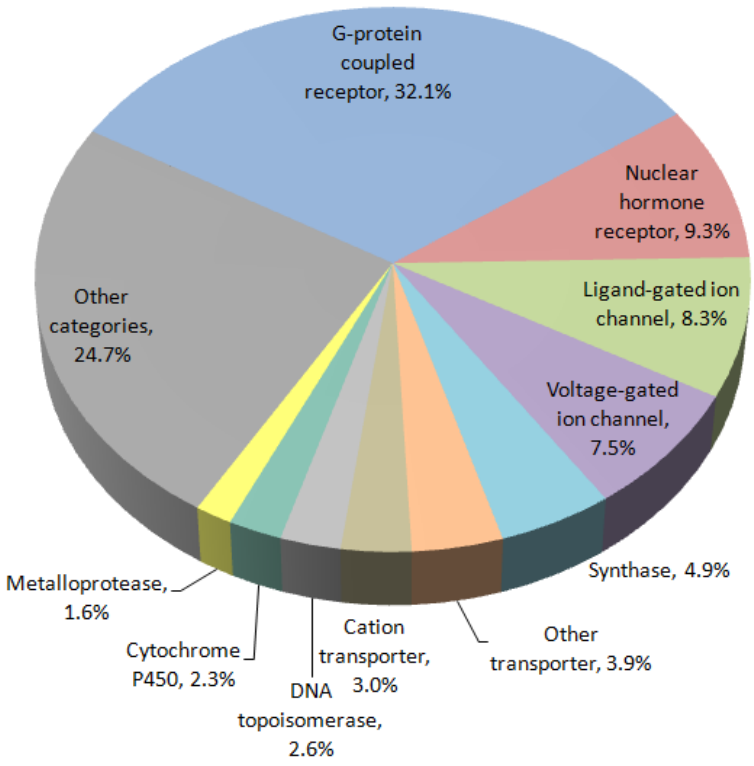


Figure 1.5 Distribution of small molecule drugs based on the targeted molecular function.

The distribution is shown for the top-ranking ten functional categories targeted by 965 FDA-approved small molecule drugs, excluding biotechnology drugs, nutraceuticals such as vitamins and amino acids, and those with uncertain targets. The top ten categories shown in the pie chart are associated with more than 75% of the drugs in the dataset. The set includes 1008 drug-protein associations. A given category is counted once if a given drug targets multiple proteins in that category. Figure is adopted from (Bahar et al., 2009).

The membrane proteins that are most frequently targeted by small molecule drugs are histamine H1 receptors, α1-adrenergic receptors and D2 dopamine receptors. All three are members of the GPCR family of proteins. These are succeeded by γ-aminobutyric-acid (GABA) A receptor α1, a LGIC. These proteins are still being investigated in relation to a broad spectrum of diseases including central nervous system disorders, allergies, inflammation, respiratory disorders, headache and sleep disorders (Zheng et al., 2006a).

Notably, most of the drugs currently in use were not initially developed to interact with a specific target protein, but to induce certain phenotypes in cultured cell or animal assays (Filmore, 2009). The identification of the targets followed the completion of the drug discovery process (a trial-and-error process using combinatorial chemistry rules). The importance of assessing drug targets and understanding the mechanistic aspects of drug-target associations became clear only in recent years. Knowledge of structure and dynamics of target proteins is now recognized to be a crucial element in making progress in rational drug discovery (Congreve et al., 2005). We think that simple models and methods for sampling alternative protein conformations may be beneficial to computational discovery efforts targeting membrane proteins. In this work, we also illustrate the utility of ensemble analysis and physics-based methods by way of calculations performed for two widely studied receptors: (i) rhodopsin, a GPCR, and (ii) nicotinic acetylcholine receptor (nAChR), a LGIC.

1.5 SPECIFIC AIMS AND OUTLINE OF THE DISSERTATION

The conformational changes observed between ligand-bound and unbound structures of a protein are usually attributed to induced fit, in the absence of data supporting alternative mechanisms or a thorough analysis of protein's dynamics. In a number of recent studies, the X-ray structural (James et al., 2003), NMR (Lange et al., 2008; Gsponer et al., 2008), or theoretical data (Tobi & Bahar, 2005; Ivetac & McCammon, 2009) provided evidence for the existence of a

conformational selection (or pre-existing equilibrium) mechanism in complex formation, in particular for protein-protein interactions. Yet, it should be noted that in these studies, the role of induced-fit was not necessarily ruled out, but was shown to follow conformational selection, rather than drive it, and be rather localized.

The interactions of proteins with small-molecule drug are even more difficult to interpret . For many widely studied targets, including those we have studied, the conformational changes observed in the ligand-bound structure are accepted to be ligand-induced, regardless of their magnitude (Cavasotto & Abagyan, 2004; Ragno et al., 2005; Hyeon et al., 2009).

In the present study, we hypothesized that the same mechanisms that govern the dynamics of protein-protein interactions also apply to protein-drug interactions; and that a small-molecule drug can induce local changes only after selecting a protein conformation that is pre-disposed to binding. These local changes may be side-chain rotations or re-orientation of binding loops. Those beyond the ability of drugs to affect are the large-magnitude collective motions of the protein, such as those engaging entire domains. A thorough comparative analysis of the growing data in the Protein Data Bank (Berman et al., 2000) shall enable us to elucidate the contribution of these two complementary mechanisms.

As detailed later, we systematically searched for targets with multiple diverse structures and performed a comparative analysis. On the practical side, we aimed at incorporating protein backbone flexibility into modeling protein-ligand interactions in order to account for the conformational effects that are usually omitted in docking simulations. To this aim, we sampled

multiple protein backbone configurations prior to docking, and then generated bound-ligand conformations for each protein conformer by considering the flexibilities of the ligand and amino acid side chains. On the biological side, we aimed at understanding and modeling MKP interactions with their inhibitors.

Our studies have been reported in five publications, consisting of three research and two review articles. We published our results from comparative study of target protein structural ensembles and intrinsic dynamics of proteins in:

- Bakan, A. & Bahar, I. (2009). The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U.S.A*, 106, 14349-14354.

We discuss these results, and the methods and tools of this research in greater detail in Sections 2 and 3.

The results from our studies of MKP inhibitors have been published in three articles:

- Lazo, J. S., Skoko, J. J., Werner, S., Mitasev, B., Bakan, A., Koizumi, F., Yellow-Duke A., Bahar I., Brummond K.M. (2007). Structurally unique inhibitors of human mitogen-activated protein kinase phosphatase-1 identified in a pyrrole carboxamide library. *J Pharmacol. Exp. Ther., 322,* 940-947.
- Bakan, A., Lazo, J. S., Wipf, P., Brummond, K. M., & Bahar, I. (2008). Toward a molecular understanding of the interaction of dual specificity phosphatases with substrates: insights from structure-based modeling and high throughput screening. *Curr. Med. Chem., 15,* 2536-2544.

- Molina, G.*, Vogt, A.*, Bakan, A.*, Dai, W., Queiroz de, O. P., Znosko, W., Smithgall T.E., Bahar I., Lazo J.S., Day B.W., Tsang M. (2009). Zebrafish chemical screening reveals an inhibitor of Dusp6 that expands cardiac cell lineages. *Nat. Chem. Biol., 5,* 680-687. *(*: equal contribution*)

We discuss these results and methods in Sections 2 and 4.

Finally, we reviewed the methods we used and their applications to membrane proteins in a *Chemical Reviews* article:

- Bahar, I., Lezon, T. R.*, Bakan, A.*, & Shrivastava, I. H. (2009). Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chem. Rev.* Published online DOI: 10.1021/cr900095e *(*: equal contribution*)

We briefly present some results for two membrane proteins in Section 5.

While we fulfilled most of our specific aims, new questions arose in relation to dynamics of protein-ligand interactions and to dynamics and inhibition of MKPs. We discuss these questions and our short-term plans to address them in the Discussion.

2.0 THEORY AND METHODS

In this thesis, we are primarily interested in changes in the backbone configuration of proteins targeted by drugs, which will be shortly referred as drug targets or targets. When modeling of protein and small-molecule interactions, the conformational variability of side-chains is efficiently taken into account by using rotamer libraries, such as those developed by Dunbrak (Dunbrack, Jr. & Karplus, 1993) or Lovell (Lovell et al., 2000). Yet, there are no standard means of incorporating complete backbone flexibility in, for example, molecular docking. For our purposes, a residue-level representation of proteins is sufficient to learn from structural ensembles and to model the equilibrium dynamics using physics-based models. We will apply what we learn from the analysis of structural ensembles to MKPs, and we will combine those-residue level approaches with those based on all-atom models toward making a more detailed assessment of its interactions with inhibitors.

In this section, we begin by describing two fundamental approaches that will be used for CG analysis of backbone equilibrium dynamics: principal component analysis (PCA) and anisotropic network model (ANM) analysis. The former is applied to the analysis of ensembles of experimentally resolved structures. The latter is a predictive model. The two approaches therefore provide information based on experiments and theory, respectively. Second, we will

show how to compare the protein motions predicted by these two methods. Third, we describe all-atom models and methods (e.g. force fields and docking) that will be used in this work, and describe how we combine low-resolution and all-atom models. Fourth, we describe supporting methods, such as surface electrostatic potential calculations which will be used in assessing druggable sites.

2.1 PRINCIPAL COMPONENT ANALYSIS (PCA) OF STRUCTURAL ENSEMBLES

Principal component analysis is an orthogonal linear transformation that transforms data from the Cartesian coordinate system into a new system of collective coordinates (Jolliffe, 2002). Its application to ensembles of structures aims at gaining a simplified view of the structural variability in the examined dataset by identifying the directions of dominant structural changes. The new coordinate system is such that the maximum possible variance in the dataset lies along the first principal component (PC) axis, and then along the 2nd PC axis, and so on.

Here, we focus on the application of PCA methods to extract information on equilibrium dynamics. PCA is performed in this case for an ensemble of structures hosted in the PDB for the same protein determined in the presence of different ligands. Alternatively, ensembles of NMR models or simulation snapshots may be used (Yang et al., 2008; Yang et al., 2009). The utility of PCA for understanding functional dynamics is clearly demonstrated by a recent study of ubiquitin structures in which a single mode of motion was found to largely account for the ability of ubiquitin to recognize a range of substrates (Lange et al., 2008).

2.1.1 Iterative optimal superimposition of structures

The first step before performing PCA is to transform (rotate and translate) all structures in the ensemble so that they are all expressed in the same reference coordinate system and their differences in the external degrees of freedom (rigid-body translation and rotation) are eliminated. This type of transformation will be called the optimal superimposition.

Let us consider an ensemble of M conformations, for a protein of N residues. Each conformation, k, is described by a $3N$-dimensional vector consisting of the position vectors $\boldsymbol{R}_i^{(k)} = (x_i^{(k)} \;\; y_i^{(k)} \;\; z_i^{(k)})^T$ of the N α-carbons ($1 \le i \le N$) in that particular conformation, organized as

$$\mathbf{q}^{(k)} = ((\boldsymbol{R}_1^{(k)})^T, \ldots., (\boldsymbol{R}_N^{(k)})^T)^{\mathbf{T}} = (x_1^{(k)}, y_1^{(k)}, \ldots, x_N^{(k)}, y_N^{(k)}, z_N^{(k)})^{\mathbf{T}} \qquad (2.1.1)$$

Note that superscript T stands for transpose, and $\mathbf{q}^{(k)}$ is a column vector.

Similarly, we define a $3N$-dimensional fluctuation vector $\Delta\mathbf{q}^{(k)} = \mathbf{q}^{(k)} - \mathbf{q}^0$ for each conformation, describing the departure $\Delta\boldsymbol{R}_i^{(k)} = \boldsymbol{R}_i^{(k)} - \boldsymbol{R}_i^0$ in the position vectors of the N sites from their equilibrium position $\boldsymbol{R}_i^0 = (x_i^0 \;\; y_i^0 \;\; z_i^0)^T$ (**Figure 2.1**). The equilibrium positions are identified by the average coordinates over all optimally superimposed PDB structures. The optimal superimposition is an iterative procedure and consists of the following steps: (i) Each structure in the ensemble is first pairwise superimposed onto a randomly selected reference structure using the Kabsch algorithm (Kabsch, 1976). This algorithm minimizes the root-mean-

squared deviation (RMSD) between a given pair of structures. Essentially, the Kabsch algorithm finds the rotation matrix for the superimposed structure that will minimize the RMSD, after the center of mass (CoM) of the superimposed structure is moved to the position of CoM of the reference structure. (ii) An average set of coordinates is calculated for the superimposed set obtained in (i), referred to as the 'average model', (iii) all structures are pairwise superimposed on the newly generated 'average model' using the Kabsch algorithm, (iv) steps (ii)-(iii) are repeated until the average model generated in two successive iterations changes by less than the threshold RMSD of 0.001Å. This procedure usually converges in 3-5 iterations. The present superposition method ensures that the structures do not undergo rigid body translational and rotational motions, and allow for direct comparison of the deformation vectors with eigenvectors from a NMA that describe purely internal motions. This iterative approach is implemented using Python scripting language and Molecular Mechanics Toolkit, as described below.

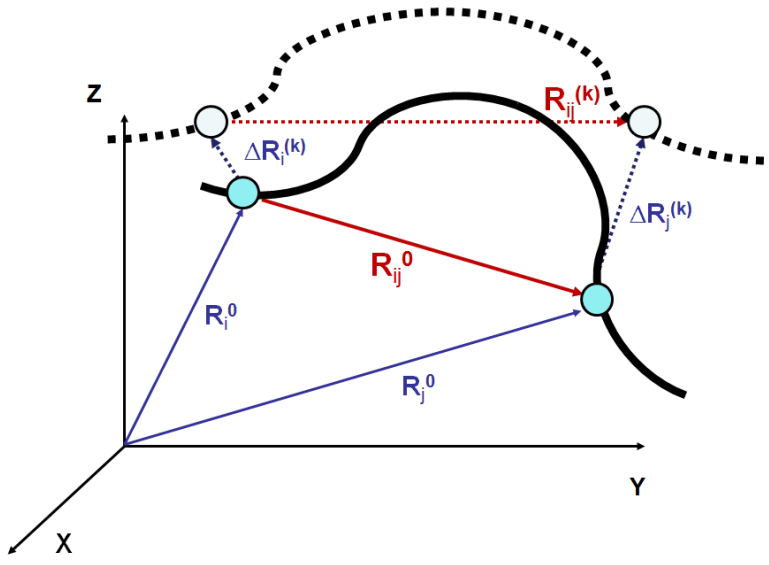


Figure 2.1 Schematic representation of a protein configuration.

In the equilibrium conformation, the positions of the α-carbons i and j are designated respectively as $\boldsymbol{R}_i^0$ and $\boldsymbol{R}_j^0$, and the vector $\boldsymbol{R}_{ij}^0 = \boldsymbol{R}_j^0 - \boldsymbol{R}_i^0$ defines the distance vector between these sites. In the k^{th} conformation in the ensembles,

atoms move to $\boldsymbol{R}_i^0 + \Delta\boldsymbol{R}_i^{(k)}$ and $\boldsymbol{R}_j^0 + \Delta\boldsymbol{R}_j^{(k)}$, and the distance vector becomes $\boldsymbol{R}_{ij}^{(k)}$. The solid black curve represents the structural details of the initial-state of the protein backbone, and the broken black curve indicates its state after a change in configuration. Figure is adapted from (Bahar et al., 2009).

2.1.2 Covariance matrix: a measure of correlations between residue fluctuations

It is of interest to understand the type and strength of coupling between the variations in different degrees of freedom. The cross-correlations between the components of the fluctuation vectors are given by the averages $<\Delta q_i \, \Delta q_j>$ over all conformations. These values can be organized in a *3N x 3N* covariance matrix $\boldsymbol{C}$,

$$\boldsymbol{C} = (1/\mathrm{M}) \, \Sigma_k \, \Delta\mathbf{q}^{(k)} \, \Delta\mathbf{q}^{(k)\mathbf{T}} \qquad (2.1.2)$$

where $\Delta\mathbf{q}^{(k)}$ denotes the 3*N*-dimensional fluctuation/displacement vector for a protein of *N* residues.

The covariance matrix $\boldsymbol{C}$ provides a detailed description of equilibrium motions, including the mean-square fluctuations in the individual interaction sites and their cross-correlations. The elements of $\boldsymbol{C}$ may be organized in submatrices of size *3x3* as

$$\boldsymbol{C_{ij}} = \begin{bmatrix} <(\Delta x_i \Delta x_j)> & <(\Delta x_i \Delta y_j)> & <(\Delta x_i \Delta z_j)> \\ <(\Delta y_i \Delta x_j)> & <(\Delta y_i \Delta y_j)> & <(\Delta y_i \Delta z_j)> \\ <(\Delta z_i \Delta x_j)> & <(\Delta z_i \Delta y_j)> & <(\Delta z_i \Delta z_j)> \end{bmatrix} \qquad (2.1.3)$$

where i and j vary in the range [1, N], both. $\boldsymbol{C}$ may thus be viewed as an NxN super matrix, the super elements of which are defined by Equation 2.1.3.

Here, we use boldface subscripts to designate a (sub)matrix or vector, and lightface subscripts for scalars (e.g., elements of vectors or matrices). The sum of the diagonal elements of $\boldsymbol{C}_{ij}$,

$$tr\{\boldsymbol{C}_{ij}\} = <\Delta x_i \Delta x_j> + <\Delta y_i \Delta y_j> + <\Delta z_i \Delta z_j> = <\Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j> \qquad (2.1.4)$$

provides a measure of the cross-correlation between the fluctuations $\Delta \boldsymbol{R}_i$ and $\Delta \boldsymbol{R}_j$ of sites i and j. The mean-square fluctuations in the positions of individual sites are given similarly by the trace of the diagonal submatrices, i.e., $tr\{\boldsymbol{C}_{ii}\} = <(\Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_i)> = <(\Delta \boldsymbol{R}_i)^2>$ using $i = j$ in Equation 2.1.4. In many applications, it proves useful to analyze the NxN covariance matrix, $\overline{\boldsymbol{C}}$, composed of the correlations between the fluctuation vectors $\Delta \boldsymbol{R}_i$, themselves,

$$\overline{\boldsymbol{C}} = \begin{bmatrix} <(\Delta \boldsymbol{R}_1)^2> & <\Delta \boldsymbol{R}_1 . \Delta \boldsymbol{R}_2> & .. & .. \\ <\Delta \boldsymbol{R}_2 . \Delta \boldsymbol{R}_1> & <(\Delta \boldsymbol{R}_2)^2> & .. & .. \\ .. & .. & .. & .. \\ <\Delta \boldsymbol{R}_N . \Delta \boldsymbol{R}_1> & .. & .. & <(\Delta \boldsymbol{R}_N)^2> \end{bmatrix} \qquad (2.1.5)$$

Note that the elements of $\overline{\boldsymbol{C}}$ are given by Equation 2.1.4.

The fluctuations in the Cartesian space are mapped onto the space spanned by the $3N$ (or N) principal axes upon diagonalizing the covariance matrix $\boldsymbol{C}$ as

$$\boldsymbol{C} = \boldsymbol{P}\,\mathbf{S}\,\boldsymbol{P}^{\mathbf{T}} = \sum_{k=1}^{3N} \sigma_k\, \boldsymbol{p_k}\, \boldsymbol{p_k}^{\mathbf{T}} \qquad (2.1.6)$$

In Equation 2.1.6, $\boldsymbol{P}$ is the unitary matrix of normalized displacements along the principal axes, also called the *principal modes* of structural changes, each given by a column vector $\boldsymbol{p_k}$, ($1 \leq k \leq 3N$), and $\mathbf{S}$ is the diagonal matrix of eigenvalues $\sigma_1, \sigma_2, \ldots, \sigma_N$, ordered in descending order, in this case. The eigenvalues are directly proportional to the variance along the principal axes. Therefore, the fractional contribution of displacements along $\boldsymbol{p_k}$ to the structural variability in the dataset is

$$f_k = \sigma_k / \boldsymbol{\Sigma}_i \sigma_i, \qquad (2.1.7)$$

where the summation is performed over all *3N* modes. Equation 2.1.6 permits us to assess the contribution of each mode, or subset of modes to the observed variance in the ensemble. For example, the square displacements in the position of the i^{th} residue induced by the top ranking *m* PC modes is

$$(\Delta \boldsymbol{R}_i)^2|_{1 \leq k \leq m} = tr\{[\sum_{k=1}^{m} \sigma_k\, \boldsymbol{p_k}\, \boldsymbol{p_k}^{\mathbf{T}}]_{\boldsymbol{ii}}\} \qquad (2.1.8)$$

2.1.3 Projection of conformations onto the subspace spanned by the PCs

The projection of a given conformational change $\Delta \boldsymbol{R}^{\mathrm{s}}$ onto $\boldsymbol{p_k}$ is found from

$$c_k^s = (\Delta \boldsymbol{R}^s)^{\mathbf{T}} \boldsymbol{p_k} \qquad (2.1.9)$$

In the extreme case of $\Delta \boldsymbol{R}^s$ perfectly aligned along $\boldsymbol{p_k}$, $c_k^s = ||\Delta \boldsymbol{R}^s||$, where the double bars designate the magnitude. This will be illustrated in **Figure 3.4**.

2.1.4 A note on the analysis of X-ray ensembles

In this work, covariance matrix diagonalization is the preferred method of performing PCA. Alternatively, one may perform singular value decomposition (SVD) on the *M*x3*N* matrix that contains coordinate data from the ensemble after the iterative optimal superimposition of structures (Jolliffe, 2002). Covariance method is preferred over SVD, because X-ray structures contain missing data, on usually flexible parts of the protein, such as loops and termini. Hence, the *M*x3*N* data matrix may contain zeros at positions corresponding to missing coordinates. SVD on such a data matrix may result in ill-defined principal modes due to incorrect average positions calculated for residues with missing coordinates. The covariance method enables to calculate the ensemble averages properly. The averages for each submatrix $\boldsymbol{C_{ij}}$ are taken over all existing pairs of residues. Hence, the calculated principal modes provide a robust representation of the dominant features of the essential dynamics.

In our analyses of X-ray ensembles, we included data on the residues that are resolved in 90% of the structures, omitting those portions not resolved. This resulted in considering around 90% of the target protein in each case (see Tables in Appendix A for a list of residues considered in each dataset of protein structures). When this threshold was relaxed (i.e. a larger portion of the

protein was considered), the PCA was unable to capture collective motions. The first few principal modes described, instead, the random fluctuations of termini and loop residues. Increasing the threshold (i.e., including a smaller portion of the protein) increased the agreement between experimental (PCA) and computational (ANM) data, but resulted in omitting more than 10% of the sequence. 90% threshold achieved a reasonable balance between ability to capture collective motions and preserving the structural integrity of the protein. In fact, de Groot and coworkers did exclude a similar fraction of residues from the C-terminal part of ubiquitin (6 out of 76 residues; 8%) because the random fluctuations at this region obscured the collective and functional motions (Lange et al., 2008).

2.2 A NORMAL MODE ANALYSIS BASED ON ANISOTROPIC NETWORK MODEL (ANM)

Normal mode analysis is an analytical means of evaluating the type and size of collective motions accessible to a mechanical system in a harmonic potential energy well (Cui & Bahar, 2006; Bahar et al., 2009). Its application to proteins dates back to early 1980s (Brooks & Karplus, 1983; Go et al., 1983; Naik et al., 1984; Levitt et al., 1985). However, only in the last decade has it become a tool widely used for exploring functional motions. A major reason behind its broader use is the observation that global modes elucidated by NMA bear functional significance. This feature became evident with the use of simplified models in CG NMA (Bahar & Rader, 2005).

The underlying assumption in NMA is that the effective potential between interacting pairs of atoms, or residues when the system is modeled at a CG scale, is expressed as a sum of harmonic potentials between all interacting pairs. While this assumption greatly simplifies the analysis of slow and energetically favorable motions of the system, it limits the validity of the analysis to the proximity of the initial, or the equilibrium, conformation. The validity of the harmonic approximation decreases as the system is displaced away from its equilibrium conformation. A second caveat arises when NMA does not explicitly contain bond length and bond angle constraints, especially when a residue level representation is considered. Distortions along the normal modes may violate such constraints (Song & Jernigan, 2006; Yang et al., 2007). These limit the accuracy of NMA to the immediate vicinity of the energy minimum.

NMA performed using elastic network model (ENM) representations of biomolecular systems are particularly useful for predicting large-scale motions. ENMs typically represent the protein at residue level. At this level of resolution, fine details of the energy landscape become smoothed (**Figure 1.2**). Different conformational substates, separated by low energy barriers become accessible. Such substates are usually sampled by local rearrangements such as rotations of side-chains or global (*en bloc*) movements of entire secondary structural elements or domains which involve changes in a minimal number of degrees of freedom.

2.2.1 Normal Mode Analysis (NMA)

NMA is based on the diagonalization of a $3N \times 3N$ matrix called the Hessian, **H**. We described below the construction of **H** and calculation of normal modes.

2.2.1.1 Hessian matrix

For our purposes, the physical system under consideration is a molecular system containing *N* residues, the Cartesian coordinates of which are given by Equation 2.1.1. We omit the superscript *k* here, since NMA is performed for a single structure (M = 1). Near the equilibrium conformation, $\mathbf{q}^0$, the potential energy can be expanded as a power series in **q** as

$$V(\mathbf{q}) = V(\mathbf{q}^0) + \sum_i \left(\frac{\partial V}{\partial q_i}\right)^0 (q_i - q_i^0) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)(q_j - q_j^0) + \ldots \qquad (2.2.1)$$

where the superscripts '0' indicate the equilibrium conformation. The first term is the value of the potential at the energy minimum, which may be set to zero. When the equilibrium conformation is at the minimum of the potential energy landscape, the first order derivative of the potential energy function is also zero, by definition. Hence, to a second order approximation, the potential is then a sum of pairwise potentials:

$$V(\mathbf{q}) = \frac{1}{2}\sum_{ij}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)(q_j - q_j^0)$$

$$= \frac{1}{2}\sum_{i,j}(q_i - q_i^0) H_{ij} (q_j - q_j^0) = \frac{1}{2}\Delta\mathbf{q}^{\mathrm{T}}\mathbf{H}\Delta\mathbf{q} \qquad (2.2.2)$$

where **H** is the Hessian matrix obtained from the second derivatives of the potential with respect to the components of **q** (or $\Delta\mathbf{q}$):

$$H_{ij} = \left(\frac{\partial^2 V}{\partial q_i \partial q_j} \right)^0 \tag{2.2.3}$$

In the same way as the covariance matrix ***C*** is organized, **H** may be thought of as an $N \times N$ matrix of 3×3 submatrices, each of which refers to the energetic contribution from the interaction of two residues.

Two important properties of the Hessian arise from Equation 2.2.3. First, **H** is real and symmetric by definition, and is therefore diagonalized by an orthogonal transformation. Second, none of the eigenvalues of **H** can be negative if **H** is constructed at a local potential energy minimum. The sign of a given eigenvalue indicates the local curvature of the potential along the corresponding mode directional vector, or eigenvector: Positive eigenvalues indicate local minima, and negative eigenvalues, local maxima. The local potential energy landscape for a system in a potential energy minimum will have only positive or zero curvature in all directions. Eigenvalues that are identically zero indicate conformational changes that have no effect on the system's (internal) potential energy. Typically, **H** has 6 zero eigenvalues, corresponding to the rigid-body rotations and translations of the molecule. Eigenvectors corresponding to smaller eigenvalues describe directions of deformations that are energetically favorable when compared to higher ranking eigenvectors (those with larger eigenvalues), as will be shown in the next subsection.

2.2.1.2 Normal modes

Normal modes are calculated by solving the equation of motion that account for the kinetic energy as well as the potential energy of the system. In this case, the form of the matrix that is to be diagonalized slightly changes, but this clarifies the physical interpretation of the normal modes. By considering the system to be a collection of classically behaving particles (oscillators), the equation of motion can be written as

$$\mathbf{M}\frac{d^2\Delta\mathbf{q}}{dt^2} + \mathbf{H}\Delta\mathbf{q} = 0 \qquad (2.2.4)$$

Here, the diagonal matrix **M** contains the masses of the nodes. Each mass is repeated three times, once for each of the node's three Cartesian coordinates. A solution to Equation 2.2.4 is the *3N*-dimensional vector $\mathbf{u}_k(t) = \boldsymbol{a}_k \exp\{-i\omega_k t\}$, where $\boldsymbol{a}_k$ is a complex vector containing both amplitude and phase factor, and ω_k is the frequency of the mode of motion represented by $\mathbf{u}_k(t)$. Substituting this solution into Equation 2.2.4, the equation of motion becomes

$$\mathbf{H}\mathbf{u}_k = \omega_k^2\mathbf{M}\mathbf{u}_k \qquad (2.2.5)$$

which is a generalized eigenvalue equation. The complete set of solutions $\boldsymbol{u}_k(t)$, $1 \le k \le 3N$, and the corresponding squared frequencies ω_k^2 may be organized as the respective columns of the matrix **U** and the elements $\lambda_k = \omega_k^2$ of the diagonal matrix $\boldsymbol{\Lambda}$ to rewrite the set of *3N* equations represented by Eq. 12 in compact notation as

$$\mathbf{H}\mathbf{U} = \mathbf{M}\mathbf{U}\boldsymbol{\Lambda} \qquad (2.2.6)$$

Equation 2.2.6 can be transformed into a standard eigenvalue equation $\widetilde{\mathbf{H}}\widetilde{\mathbf{U}} = \widetilde{\mathbf{U}}\mathbf{\Lambda}$ in mass-weighted coordinates through the transformations $\widetilde{\mathbf{U}} = \mathbf{M}^{1/2}\mathbf{U}$ and $\widetilde{\mathbf{H}} = \mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-1/2}$. The mass-weighted Hessian, $\widetilde{\mathbf{H}}$, retains the symmetry of the original Hessian, and its eigenvectors $\widetilde{\mathbf{u}}_\mathbf{k} = \mathbf{M}^{1/2}\mathbf{u}_\mathbf{k}$ form an orthonormal basis set (i.e., $\widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} = \mathbf{1}$). These are the *normal modes* of the system. Their Cartesian counterparts are found through the inverse transformation $\mathbf{U} = \mathbf{M}^{-1/2}\widetilde{\mathbf{U}}$ and satisfy the orthonormality condition $\mathbf{U^T M U} = 1$. If the nodes have all equal mass m, $\mathbf{M}$ reduces to the identity matrix multiplied by m, $\tilde{\mathbf{U}} = m^{1/2}\mathbf{U}$ and $\widetilde{\mathbf{H}} = m^{-1}\,\mathbf{H}$. When all nodes in the system are assumed have unit masses, energetic modes and normal modes become identical.

2.2.1.3 Interpretation of Normal Modes: Significance of slow modes

The energy associated with a given normal mode is directly proportional to the square of its frequency (or its eigenvalue $\lambda_k = \omega_k^2$). This can be seen by rewriting Equation 2.2.2 for a single mode k:

$$V(\mathbf{u}_k) = \frac{1}{2}\mathbf{u}_k^T\mathbf{H}\mathbf{u}_k = \frac{\omega_k^2}{2} \qquad (2.2.7)$$

Displacements along high-frequency modes are therefore energetically more expensive than those of equal magnitude along low-frequency modes. The vibrational energy is, on average, equally partitioned among all the modes, such that the average amplitude of oscillation along mode k scales with $1/\omega_k^2$. Thus, the molecule experiences the greatest displacement along

the lowest frequency, or 'slowest' modes, hence their qualification as 'soft modes'. These modes are also of highest interest when seeking to determine the most probable *global* fluctuations of a molecule. Large eigenvalues, on the other hand, indicate directions of steep energetic ascent, subject to high frequency, low amplitude fluctuations.

The cross-correlations $<\Delta q_i \Delta q_j>$ between the displacements of the interaction sites along different coordinates are calculated as statistical mechanical averages of the form

$$\left\langle \Delta q_i \Delta q_j \right\rangle = \frac{1}{Z} \int d^{3N-6} q e^{-\frac{1}{2k_B T} \Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} \Delta q_i \Delta q_j = k_B T \left(\mathbf{H}^{-1} \right)_{ij}, \tag{2.2.8}$$

using the configurational integral

$$Z = \int d^{3N-6} q e^{-\frac{1}{2k_B T} \Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} = \left(2\pi k_B T \right)^{3N-6/2} \left[\det \left(\mathbf{H}^{-1} \right) \right]^{1/2} \tag{2.2.9}$$

Here the integrations are performed over the complete space of equilibrium fluctuations, k_B is the Boltzmann constant, T is the absolute temperature, and $(\mathbf{H}^{-1})_{ij}$ designates the ij^{th} element of the inverse of **H**. Because of the way the elements of **H** are defined (i.e. the diagonal elements found from the negative sum of off-diagonal terms), the columns/rows are not independent, and **H** is not invertible (its determinant is equal to zero). **H** has exactly six eigenvalues that are identically zero, corresponding to the three translational and three rotational degrees of freedom. The inverse of **H** is therefore replaced by the *pseudo-inverse*, which is the inverse evaluated only in the space corresponding to the non-zero eigenvalues,

$$\tilde{\mathbf{H}}^{-1} = \sum_{k=1}^{3N-6} \frac{\tilde{\mathbf{u}}_k \tilde{\mathbf{u}}_k^T}{\omega_k^2} \qquad (2.2.10)$$

The importance of the slow modes is again highlighted in these equations: The lowest frequency modes contribute most to the spatial partition function because $\det\left(\tilde{\mathbf{H}}^{-1}\right)$ is the product of the reciprocal nonzero eigenvalues of $\tilde{\mathbf{H}}$.

2.2.1.4 Covariance computed from NMA: Bridging with PCA of structural ensembles

The cross-correlation $<\Delta q_i \Delta q_j>$ on the left-hand side of Equation 2.2.8 is simply the ij^{th} element of the covariance matrix **C**; therefore Equation 2.2.8 may be rewritten in compact notation as

$$\mathbf{C} = k_B T \, \mathbf{H}^{-1} \qquad (2.2.11)$$

This equation establishes the bridge between the PCA of ensembles of conformations and NMA of a given structure. In the former case the top-ranking (principal) modes of structural changes are extracted from experimental data (or sets of known structures for a given protein). In the latter, the same such structural changes are *predicted* by the theory using one structure to construct **H**.

The top-ranking modes obtained by PCA should, in principle, be comparable to the lowest frequency modes derived by NMA (i.e. $\lambda_i \sim 1/\sigma_i$, and $\boldsymbol{p_i} \sim \boldsymbol{u_i}$), provided that (i) the dataset of conformations subjected to PCA represents an equilibrium distribution, and (ii) the Hessian in NMA provides an accurate description of dominant interactions. Recent PCAs performed for

ensembles of PDB structures exhibit good agreement with the global modes predicted by CG NMAs (Yang et al., 2008; Yang et al., 2009). Notably, ENMs have been adopted in those NMAs. The relevance of ENM predictions for a given protein to PC modes derived from sets of structures experimentally resolved for the same protein (under different conditions, in the presence of different ligands) lends support to the use of ENMs for assessing functional changes in structure.

2.2.2 ANM: Theory and Foundations

The ANM, owing to its simplicity, is one of the most commonly used ENMs (Doruker et al., 2000; Atilgan et al., 2001; Tama & Sanejouand, 2001; Eyal et al., 2006). It approximates the protein's potential energy as that of a classical network of identical masses coupled by uniform springs: each node in the network may correspond to an atom or a residue, and each edge is a spring. This approach accepts a PDB structure as the global energy minimum; hence the computationally expensive energy minimization step is not necessary. In the ANM, the network topology is defined by the native structure, with edges placed between nodes within a pre-specified cutoff distance.

2.2.2.1 Assumptions and Model Parameters

Comparisons of predicted root-mean-squared (RMS) fluctuations to motions inferred from crystallographic B-factors have identified optimal cutoff distance of 18Å for the ANM (Eyal et al., 2006). As to the spring constants, the simplest ENMs use a uniform force constant for all

interactions. Hinsen proposed using a force constant that decays rapidly with distance (Hinsen, 1998). Sen and Jernigan empirically investigated how the force constant should vary with the coordination number of residues (Sen & Jernigan, 2006). The adoption of stiffer springs for sequentially neighboring residues or amino acid-specific force constants has been shown to improve the agreement with experiments.

The choice of the specific spring constants has little, if any, effect on the global modes, which are the primary focus of this work. The global modes of motion are widely recognized to be intrinsic properties of the 3D shape of the protein, and have been verified in several studies to be insensitive to model parameters (Doruker et al., 2002; Ma, 2005; Lu & Ma, 2005; Nicolay & Sanejouand, 2006; Zheng et al., 2006b) and almost identically reproduced at various hierarchical levels of resolutions (Doruker et al., 2002; Chennubhotla & Bahar, 2006; Chennubhotla et al., 2008). This robustness of global modes permits us to utilize the ANM as shown below.

2.2.2.2 Potential Function and the Hessian

ANM analysis is simply a NMA subject to the potential (Hinsen, 1998)

$$V = \frac{1}{2}\sum_{ij}\gamma_{ij}\left(R_{ij} - R_{ij}^{0}\right)^2 \qquad (2.2.12)$$

Using this expression in Equation 2.2.3, it is possible to readily write a closed form expression for **H** as

$$\frac{\partial^2 V}{\partial x_i \partial y_j} = -\frac{\gamma_{ij}\left(x_j - x_i\right)\left(y_j - y_i\right)}{R_{ij}^2} \tag{2.2.13}$$

Using the notation $x_{ij}^0 = \left(x_j^0 - x_i^0\right)$ and similar expressions for the three components of the instantaneous distance vector $\boldsymbol{R}^0_{ij}$, the off-diagonal 3x3 submatrices of **H** take the form

$$\mathbf{H}_{ij} = -\frac{\gamma_{ij}}{\left(R_{ij}^0\right)^2}\begin{bmatrix} \left(x_{ij}^0\right)^2 & x_{ij}^0 y_{ij}^0 & x_{ij}^0 z_{ij}^0 \\ x_{ij}^0 y_{ij}^0 & \left(y_{ij}^0\right)^2 & y_{ij}^0 z_{ij}^0 \\ x_{ij}^0 z_{ij}^0 & y_{ij}^0 z_{ij}^0 & \left(z_{ij}^0\right)^2 \end{bmatrix} \tag{2.2.14}$$

and the diagonal submatrices satisfy the identity

$$\mathbf{H}_{ii} = -\sum_{j;j\neq i} \mathbf{H}_{ij} \tag{2.2.15}$$

The above simple expressions for the elements of **H** are readily used in NMA to determine the collective dynamics.

We note that amino acid specificity can be included in ENM-based studies by adopting residue-specific force constants, γ_{ij}, that are dependent on the identity of the amino acids *i* and *j* connected by a spring in the network. However, in most applications, γ_{ij} is taken as a constant, γ, for all pairs of residues connected in the network. Equation 2.2.12 with a single parameter $\gamma_{ij} = \gamma$ has been originally used by Tirion for representing *interatomic* interactions (as opposed to *inter-residue* interactions considered in all succeeding ENM studies, starting from the Gaussian

Network Model (Haliloglu et al., 1997; Bahar et al., 1997) based on the statistical thermodynamics of polymer networks, and other CG ENMs such as the ANM. Several studies demonstrated to date the ability of these CG models to almost identically reproduce the global modes obtained by atomic NMA that use detailed force fields. As mentioned above, the absolute value of γ for a given level of representation does not affect the mode shapes (i.e., the eigenvectors, $\boldsymbol{u}_k$, ($1 \leq k \leq 3N$-6) of **H**), but their frequencies, because the eigenvectors represent distributions of mobilities, irrespective of the eigenvalues. The eigenvalues, on the other hand, are proportional to γ. In other words, the same distribution of motions, or relative mobilities is obtained, irrespective of the choice of g, which uniformly scales the absolute sizes of all residues' motions. Previous studies also showed that the global modes are practically insensitive to specific inter-residue interactions, or the adoption of residue-specific force constants. The most important determinant of global modes is the connectivity of the matrix, itself, regardless of the weight (or stiffness) of the edges (or springs) that connect the nodes (residues).

2.2.2.3 Generation of Alternative Conformations Using the ANM

A major utility of the ANM is its ability to generate alternative conformations (substates or microstates) in the close neighborhood of a given structure. These are generated upon deforming the original structures along the softest modes determined by ANM analysis. Similar to Equation 2.1.8, the square displacement of residue i contributed by the movement along a given mode k is given in terms of the k^{th} eigenvector ($\boldsymbol{u}_k$) and eigenvalue (λ_k) of **H** as

$$(\Delta \boldsymbol{R}_i)^2|_k = tr\{[\lambda_k^{-1}\, \boldsymbol{u_k}\, \boldsymbol{u_k}^{\mathbf{T}}]_{ii}\} \qquad (2.2.16)$$

Or the alternative conformations induced upon moving along a given mode are simply,

$$\mathbf{q}^{(k)} = ((\boldsymbol{R}_1^{(k)})^T, \ldots., (\boldsymbol{R}_N^{(k)})^T)^{\mathbf{T}} = ((\boldsymbol{R}_1^{0})^T, \ldots., (\boldsymbol{R}_N^{0})^T)^{\mathbf{T}} \pm s\, \lambda_k^{-1/2} \boldsymbol{u_k} \qquad (2.2.17)$$

where the coefficient s scales with $(k_B T)^{1/2}$. In principle, given the uncertainty in the absolute value of γ, which is reflected on the eigenvalues, a range of s values giving rise to movements comparable in size to those experimentally observed is generated, and used for further calculations, such as generating an ensemble of conformations to be used in docking simulations.

2.3 COMPARISON OF ESSENTIAL MODES DERIVED FROM PCA AND ANM ANALYSIS

When the structure of a protein is known in alternative forms, one can use that information to select effective normal modes to achieve a given transition. T contribution of these modes, i.e., the choice of s in Equation 2.2.17, is usually based on the correlation cosine, or *overlap*

$$O_k = (\Delta \boldsymbol{q}_{AB} \cdot \boldsymbol{u_k})/|\, \Delta \boldsymbol{q}_{AB}|, \qquad (2.3.1)$$

between the normalized directional vector $\boldsymbol{u_k}$ and the targeted direction of deformation $\Delta \boldsymbol{q}_{AB} = \mathbf{q}^{(B)} - \mathbf{q}^{(A)}$ provided that the goal is to explore the transition from substate A to B (Marques &

Sanejouand, 1995). Similarly, the overlap between the PCA mode *i* and the ANM mode *j* may be calculated as

$$O_{ij} = \boldsymbol{p_i} \cdot \boldsymbol{u_j} \tag{2.3.2}$$

Note that ANM modes define a complete space of conformational changes, i.e., each eigenvector may be viewed as a component/axis of a complete 3N-6 dimensional orthonormal basis set. As a result, the summation over all modes satisfy the relation,

$$\sum_{j=1}^{3N-6} O_{ij} = 1 \tag{2.3.3}$$

In the following chapter, when we compare the PCA modes derived from experiments with the ANM-predicted normal modes, we will seek a correspondence between the top ranking 3 PCA modes and softest 20 ANM modes. It will be seen that one of the three softest 3 ANM modes will exhibit the closest correspondence to top-ranking experimental modes. In some cases, the combination of two slow modes will be shown to reach a significant level of correlation (correlation coefficient around or above 0.7) with experimental data (e.g. the case of calmodulin discussed in the next chapter).

In the comparison of PCA and ANM, another quantity of interest is the cumulative overlap that describes the contribution of subsets of ANM modes to a PCA mode

$$CO_i^J = \left[\sum_{j=1}^{J} \left(O_{ij} \right)^2 \right]^{1/2} \tag{2.3.4}$$

where the summation is performed over the subset of ANM modes of interest, usually starting from the lowest-lying modes. As explained above, this summation is identically equal to unity if it is performed over all 3*N*-6 modes/eigenvectors.

Finally, the subspace overlap between the PCA and ANM essential spaces spanned by *I* PCA and *J* ANM modes, respectively, is found from (Amadei et al., 1999)

$$SO^{IJ} = \left[\frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(O_{ij} \right)^2 \right]^{1/2} \tag{2.3.5}$$

2.4 ALL-ATOM MODELS OF PROTEINS

A residue-level representation of the protein structure is adopted in ANM and PCA calculations as explained above. However, understanding the interactions of proteins with small-molecules requires consideration of chemical specificity at atomic scale. Hence, an all-atom representation of the protein structure and energetics is an inevitable ingredient of our modeling studies. In all-atom representations of proteins, or in general of chemical entities, electrons are generally ignored and each atom is represented by the position of its nucleus with a mass and partial or full charge (Leach, 2001). This is the most common representation of molecular structures in MD

simulations and in modeling protein-ligand interactions. Finer representations with full or partial electronic detail such as quantum mechanical or semi-empirical methods are only applicable to systems with small number of atoms, in the orders of ten, due to the substantially increased computational cost.

2.4.1 Force Fields

All-atom representations of molecules are commonly and conveniently described by molecular mechanics (MM) force fields (FFs). FFs contain multiple atom type definitions for each element to capture the differences in its chemical environment in different molecules (Leach, 2001). For example, oxygen atoms in ether and ester groups are described using different atom types due to the differences in their bonded atoms and bond orders. FFs generally limit the number of atom types per element, which makes it feasible to parameterize the potentials between chemically bonded or physically interacting atom pairs. FF parameters are derived from quantum mechanical calculations of small representative compounds and are fitted to experimental spectral and/or thermodynamic data, hence are also called empirical force fields. A typical potential function is composed of additive components describing bonded and non-bonded interactions (Leach, 2001; Wang et al., 2004; Brooks et al., 2009). The total potential energy of the system is written as the sum of bonded and non-bonded potential energies:

$$V_{FF} = V_{bonded} + V_{non\text{-}bonded} \tag{2.4.1}$$

Bond length and bond angle potentials are defined as quadratic functions around their equilibrium values. Bond rotational, or dihedral angle, potentials, on the other hand, involve four bonded atoms, and are conveniently defined as a sinusoidal function. The contribution of bonded interactions to the total energy may be written as:

$$V_{bonded} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\psi(1 + \cos(n\psi - \varphi)) \qquad (2.4.2)$$

Here, K_b, K_θ, and K_ψ are the force constants with units in kcal/mole/Å^2, kcal/mole/rad^2, and kcal/mole, respectively; b_0 and θ_0 are the equilibrium bond length and bond angle values. ψ demontes the dihedral angle. n is the number of minima on the potential energy function of the dihedral angle which defines the periodicity of the bond rotational potential (e.g., $n = 3$ for sp^3 C-C bonds) , and φ is the phase lag parameter, to set the location of these minima. These parameters are defined for all possible groups of atoms that may exist in the simulated system.

Non-bonded potentials are usually expressed as a sum over van der Waals (vdW) potentials (e.g. expressed in terms of Lennard-Jones function), and Coulombic interactions:

$$V_{non\text{-}bonded} = V_{Lennard\text{-}Jones} + V_{Coulombic} \qquad (2.4.3)$$

The vdW forces describe the attractive and/or repulsive forces between atoms arising from permanent or induced dipoles. The Lennard-Jones (LJ) potential, commonly used to describe these forces, is expressed in terms of the interatomic distance r_{ij} as:

$$V_{Lennard\text{-}Jones} = \sum_{\substack{\text{non-bonded} \\ \text{atom pairs}}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.4.4}$$

where hard-sphere diameter (σ_{ij}) and potential well depth parameter (ε_{ij}) depend on the pair of interacting atoms, and the summation includes all pairs i and j separated by at least three intervening bonds. The first term in the LJ potential models the strong repulsion between two atoms that arise when their electron clouds overlap. The second accounts for the attractive interactions. The equation has its minimum value at $2^{1/6}\sigma_{ij}$, and quickly decays to zero (e.g., at r_{ij} < 10Å) with increasing interatomic distance. For computational efficiency, the LJ potentials are generally evaluated for pairs of atoms within a certain cutoff distance (8-12Å). Also, it is customary to adopt two different sets of parameters, distinguishing between near neighbors (separated by 3 bonds) and farther neighbors along the chain.

Differences in the electronegativity of atoms making up an organic molecule leads to uneven distribution of charge on the molecule, even when the molecule is neutral. Classical FFs assign a point charge to each atom based on atoms that are chemically bonded to it. These charges do not change during a simulation due to varying physical environment. Columbic potential is used to calculate the energy that arises from these charged interactions:

$$V_{Coulombic} = \sum_{\substack{\text{non-bonded} \\ \text{atom pairs}}} K_{Coulombic}\frac{q_i q_j}{\varepsilon_D r_{ij}} \tag{2.4.5}$$

where q_i and q_j are partial charges assigned to atoms i and j in fundamental unit of electronic charge (e), ε_D is dielectric constant, whose value may be chosen to depend on distance or medium (e.g. 80 for bulk water), and the proportionality constant $K_{Coulombic}$ is 322 kcal*Å/(mol*e^2) to express the energy in kcal/mol.

In this work three different molecular mechanics FFs are used. A specific FF for a particular task is chosen based on the applicability of the force field to that task. When multiple force fields are applicable for a certain task, the one that is consistent with prior works of others is preferred.

AMBER

AMBER (Assisted Model Building with Energy Refinement) is a popular MM FF initially developed for performing MD simulations of proteins and nucleic acids (Wang et al., 2004; Yang et al., 2006). This FF is part of the AMBER simulation package. It is also implemented in many molecular modeling packages such as Sybyl. We used this FF for geometry optimization, or energy minimization, of protein structures prior to their use in small-molecule docking simulations. The reason that this FF is used is that its parameters and partial charges are used in development of molecular docking software, such as AutoDock (Huey et al., 2007) and DOCK (Meng et al., 1992). Therefore, optimizing the protein structure and assigning partial charges using AMBER should result in more consistent scoring of the interactions of proteins with small-molecules.

CHARMM

CHARMM (Chemistry at HARvard Macromolecular Mechanics) is another MM FF designed for MD simulations of proteins and nucleic acids (Brooks et al., 2009). It is implemented for convenient use in NAMD (a molecular simulation package described below). We used CHARMM implementation in NAMD in combination with ANM modes to generate alternative structures of target proteins with proper internal geometry.

MMFF94

MMFF94 (Merck molecular force field) is a FF developed for studies of small organic compounds (Halgren & Nachbar, 1996; Halgren, 1996a; Halgren, 1996b; Halgren, 1996c; Halgren, 1996d; Bush et al., 1999; Halgren, 1999a; Halgren, 1999b). It is implemented in many computer aided drug-discovery oriented software packages, such as Sybyl (described below). We used MMFF94 to prepare and optimize all small-molecule ligand structures prior to docking simulations.

2.4.2 Molecular modeling and simulation packages

The protein structures deposited in the PDB contain coordinates of almost all heavy (other than hydrogen) atoms, which is sufficient for many molecular modeling purposes. The standard procedure in modeling protein-ligand interactions includes adding hydrogen atoms to raw PDB structures; placing missing atoms of side-chains or flexible loops of the protein, adding atomic partial charges, and optimization of the atomic coordinates based on a force field. We utilized the following modeling and simulation packages and their force fields:

Sybyl

Sybyl is a commercial industry-standard molecular modeling program from Tripos (http://www.tripos.com/) that has determined many standards in representation of small-molecules in computer media. We used this software to prepare and energy minimize structures of proteins and small-molecules. Protein energy minimization was essentially used to optimize the positions of the newly added hydrogen atoms, so heavy protein atom positions were fixed. All minimizations were performed using default method and parameters, except the termination criterion. The minimization was terminated when an energy gradient of 0.05 kcal/mol/Å was reached.

NAMD

NAMD is a scalable MD simulation package (Phillips et al., 2005). NAMD can incorporate external forces, forces that are not defined in the FF, in minimization or simulation protocols in terms of harmonic restraints on a subset of atoms. This enables deforming a structure along global modes. We used NAMD to minimize a protein structure using CHARMM force field and ANM restraints. Details of this procedure will be described below.

2.4.3 Comparative modeling

When the structure of the protein of interest is not available, but structure(s) of sequentially homologous protein(s) are known, comparative modeling techniques may be used to obtain the unknown structure based on the known structure(s) and the sequence alignment of the target and its template(s). Comparative modeling uses constraints based on the template structures in

addition to those from all-atom FFs, rotamer libraries and databases of protein structures. The template structures provide constraints on the fold and secondary structural elements of the target protein. FFs and rotamer libraries assist in the placement and optimization of side-chains.

In this work, we used two different comparative modeling programs:

MODELLER

MODELLER is an academically available comparative modeling software (Sali & Blundell, 1993). Given sequence alignment and template structures, MODELLER derives spatial restraints and expresses them as probability density functions. The target model is then determined using a variant of the structure determination algorithm developed by Braun and Go (Braun & Go, 1985) which works on the Cartesian space. Due to the random nature of the search method, we generated multiple models for the target. When we needed a single conformation of the target protein, we selected the model with the highest (best) probability score. In cases where we needed multiple target conformations, we generated several hundreds of models and used a subset with the highest scores.

ORCHESTRAR

ORCHESTRAR is commercially available comparative modeling software (part of Sybyl from Tripos, http://www.tripos.com/) that works on the same basic principles. The advantage of ORCHESTRAR is that the user has control over the entire process. The modeling process is completely visualized and performed in steps: (i) core backbone structure is modeled, (ii) loops are modeled, (iii) side-chains are modeled, (iv) hydrogen atoms are added and structure is energy minimized. We used this program when we needed a single model of the target structure.

2.5 ALTERNATIVE PROTEIN CONFORMATIONS FROM NORMAL MODE GUIDED DISPLACEMENTS

Use of alternative protein conformations when docking small molecule ligands improve the results (Totrov & Abagyan, 2008). One can find a diverse set of conformations in the PDB for drug targets only in a few cases. The NMA becomes useful when such conformations are needed. It helps guide the search on the potential energy surface of the target protein by enumerating a set of conformations accessible via soft modes, which may be relevant to ligand binding (Cavasotto et al., 2005; Floquet et al., 2006; May & Zacharias, 2008).

In this work, we used the ANM modes with all-atom energy minimization to generate alternative conformations for our targets. NAMD was used for energy minimization. The implementation of the procedure was analogous to that of ANM restrained MD simulations, where a set of harmonic restraints steers the simulation to a target conformation that lies at the minimum of the harmonic restraints (Isin et al., 2008). In our case, ANM-based harmonic restraints (ANM-HR) were incorporated in energy minimization scheme, using the expression

$$V_{ANM-HR} = \sum_{i}^{N} K_{ANM-HR}\left[\boldsymbol{R}_i - \left(\boldsymbol{R}_i^0 + s\boldsymbol{v_i}\right)\right]^2 \qquad (2.5.1)$$

Here, the summation is performed for all α-carbons. $\boldsymbol{R}_i^0$ is the starting position for the i^{th} α-carbon. $\boldsymbol{v}$ is a single or a combined ANM mode. $\boldsymbol{v_i}$ is the 3-dimensional component of this mode corresponding to the i^{th} α-carbon. $\boldsymbol{v}$ may be selected to be the ANM mode that yields highest overlap with a known structural deformation vector according to Equation 2.3.1 or with a PCA mode according to equation 2.3.2. It may also be any vector that describes a relevant change in structure, such as the fluctuations of a binding site loop. If there are multiple relevant ANM modes, they may be linearly combined using the coefficients from Equations 2.3.1 or 2.3.2 to yield a combined mode. *s* is a scaling factor, to set the maximum displacement for any α-carbon. The maximum displacement for an α-carbon was set to be 0.2 Å. K_{ANM} is the force constant and is set to 40 kcal/mol/Å^2. This may seem large for a single atom, but it should be noted that this restraint applies to all atoms of a residue. So, the effective force constant per atom is 5 kcal/(molÅ^2), when the average number of heavy atoms per residue is taken to be 8.

After defining the ANM restraints, the effective potential becomes

$$V = V_{FF} + V_{ANM\text{-}HR} \qquad (2.5.2)$$

For comparative purposes, the force constant for bond stretching potential between carbon atoms of CH_3-CH_3 is 222.5 kcal/mol/Å^2. The bond angle bending potential for of the bonds CH_3-CH_2-CH_3 has a force constant of 53.5 kcal/mol/rad^2 (Brooks et al., 2009). ANM restraints are therefore considerably softer than those confining bond lengths and bond angles to their equilibrium values. The minimization of protein structure subject to Equation 2.5.2 is not therefore expected to result in unrealistic distortions in bonded atom geometries.

Alternate conformations were generated using an iterative procedure, starting from the X-ray structure. At each iteration, the protein was minimized for 200 steps, using the standard conjugate gradient algorithm in NAMD. ANM calculations were performed only for the initial conformation. It is possible to recalculate ANM modes at each step (Isin et al., 2008), but as will become obvious in the results section, the slow ANM modes calculated for the unliganded reference structure drive the collective and large magnitude reconfigurations. Hence, a single ANM calculation for the initial structure of the protein is sufficient for our purposes, as these collective modes are practically insensitive to the small changes in atomic coordinates. When alternative conformations were generated for docking, an iterative procedure was adopted to generate the conformation along the two opposite directions for each ANM mode. In the application of this method to model MKP-3 general acid loop activation, an iterative procedure was used to generate conformations along the direction that closes the general acid loop.

2.6 MOLECULAR DOCKING

Protein-ligand docking can be defined as computational modeling of the binding mode of a small organic compound to a protein (Brooijmans & Kuntz, 2003). This involves two components: (i) a search algorithm that generates multiple binding poses of the small-molecule on the protein surface, and (ii) a scoring function that assigns a score to each pose based on pair-wise atomic interactions between the small-molecule and the protein. The score is, or scales, with the negative of the calculated binding free energy, and either metric, score or energy, will be adopted

for evaluating binding poses. Different categories of both search methods and scoring functions and their development have been reviewed in the literature comprehensively (Kitchen et al., 2004).

In this work, docking simulations have been performed to solve two problems: (i) locating potential binding sites for an uncharacterized ligand, and (ii) characterizing optimal interactions of a ligand at a given binding sites. Two different docking programs are used for these purposes as described below.

2.6.1 Unbiased docking simulations using AutoDock

Unbiased docking requires to search the entire protein surface for potential binding sites. This is an exhaustive search. AutoDock is the preferred software for its ability to efficiently dock flexible ligands (Goodsell & Olson, 1990; Morris et al., 1996; Morris et al., 1998; Huey et al., 2007).

AutoDock uses Lamarckian genetic algorithm (GA) as the primary search method and an empirical scoring function as the scoring method (Morris et al., 1998). The GA is a global search heuristic inspired by evolutionary biology. In the AutoDock implementation of GA, the ligand corresponds to a gene; the variables that define its state (e.g., translation, orientation, and conformation with respect to the protein) form the genotype; and the corresponding Cartesian coordinates are the phenotype. Initially, a given number of ligand states (individuals) are randomly generated. During the search, randomly selected pairs of individuals are subjected to

crossover to generate a new offspring. Also, a certain fraction of the offspring is subjected to mutation operation to introduce new traits into the population. The fitness of the individuals is determined by the scoring function (Equation 2.4.6). The size of the population is kept constant by selecting a predetermined number of individuals based on their fitness. The length of the search is determined by the maximum number of fitness evaluations or by the maximum number of generations. Multiple docking runs are in general necessary due to the stochastic and heuristic nature of the search. This search method can handle flexible molecules with up to twelve of rotatable bonds.

The AutoDock scoring function is an empirical function that predicts binding free energy and thereby the binding affinity of the ligand. It is formulated as:

$$\Delta G_{binding} = \Delta G_{vdw} \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \Delta G_{hbond} \sum_{i,j} E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right)$$

$$+ \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + \Delta G_{tor} N_{tor} + \Delta G_{sol} \sum_{i,j} \left(S_i V_j + S_j V_i\right) e^{\left(-r_{ij}^2 / 2\sigma^2\right)} \qquad (2.6.1)$$

where the ΔG coefficient preceding each term is a scalar that has been calibrated against a training set of 30 protein-ligand complexes with known X-ray structures and binding free energies. The first tem is the vdW energy term with parameters from the united atom AMBER force field (Yang et al., 2006), which implicitly models hydrogen atoms bonded to carbons. The second term is a modified LJ potential that provides an improved representation of hydrogen bonds. $E(t)$ is a directional weight based on the angle between the hydrogen bond acceptor and donor. The third term is the Coulombic potential with a distance-dependent dielectric constant

($\varepsilon(r_{ij})$). The fourth term is the ligand torsional entropy term, which adds a constant penalty to the overall score for each rotatable bond that is immobilized upon protein binding. The last term is the solvation term based on pairwise interactions of atoms with parameters based on atom types (V is volume of atoms and S is an atom specific solvation parameter) and weighted according to interatomic distance (σ = 3.5 Å) (Morris et al., 1998). These are the terms most commonly used in docking scoring functions (Kitchen et al., 2004). Protein conformational entropy is omitted.

In our unbiased simulations we used AutoDockTools, part of MGLTools (http://mgltools.scripps.edu/), to prepare input files for energy grid calculations and docking. We set the population size parameter to 250 (default 150), number of maximum fitness evaluations parameter to 5-12 million (default 2.5 million), and number of independent GA runs to 30-200 (default 10). The remaining parameters were set to the default values (Morris et al., 1998). The large number of docking poses generated independently were analyzed using clustering methods described below.

2.6.2 Focused docking simulations using GOLD

Focused (or biased) docking, on the other hand, identify optimal interactions between a ligand and a known binding site of a protein. We used GOLD docking software for this purpose (Jones et al., 1997). GOLD also uses a GA as the search method. It yields different types of scoring functions. GOLD was chosen as the focused docking software due to its ability to enable side-chain flexibility during docking. In our simulations, the solvent-accessible side-chains that were not involved in specific interactions within the protein were allowed to rotate freely. Protein

flexibility was introduced by use of multiple protein structures either coming from experiments, homology modeling or ANM calculations. Incorporating protein flexibility was particularly important for MKPs, as they contain flexible loops and many solvent exposed side-chains at their active sites.

The GOLD fitness function consists of four components: (i) protein-ligand hydrogen bond energy (external H-bond), (ii) protein-ligand vdW energy (external vdW), (iii) ligand internal vdW energy (internal vdW), and (iv) ligand torsional strain energy (internal torsion). The total GOLD fitness score is the negative sum of the above terms. Unlike AutoDock scoring function, the scoring function is not fitted to experimental data. That is, the components are not multiplied by any coefficients, so the GOLD scores are not in the same range as typical small-molecule binding free energies (-6 kcal/mol to -14 kcal/mol). Also, please note that the charged interactions are implicitly incorporated into the H-bond term (Jones et al., 1995).

When running GOLD, we set the maximum number of operations (energy evaluations) to 100,000 and GA population size to 100. Since the search space in this case is smaller when compared to that of an unbiased docking simulation, the energy evaluation and population size parameters were set relatively smaller. The remaining GA-related parameters were kept at their default values (Jones et al., 1995).

2.6.3 The need for generating large ensembles of docking poses and post-docking clustering analysis

The primary aim of docking scoring functions is to find the correct docking pose for a given ligand from amongst many alternative poses at its known binding site. These functions work best when the protein conformation resembles the bound conformation and the search is limited to the known binding site (Totrov & Abagyan, 2008). Most docking algorithms and scoring functions do not account for protein conformational and entropic effects. It is the task of the user to generate multiple protein conformations that can potentially bind the ligand and to incorporate entropic effects when evaluating docking poses.

Recently, Ruvinksy showed that the energy well associated with the experimentally observed binding mode of ligands is broader than all energy wells associated with non-native binding poses (Ruvinsky, 2007). In other words, the bound conformation accommodates slight variations in the binding pose without affecting (weakening) protein-ligand interaction energy (enthalpy), and this variability, or associated favorable entropy, helps in lowering the free energy of binding.

An *ensemble modeling* approach that takes into account the width of the energy wells was recently adopted to explain the differences in the catalytic rates of cytochrome P450 orthologs that process the same substrates (Prasad et al., 2007). These studies provide us with insights into how one should select the most likely docking poses. Using molecular docking programs, an improved accuracy of predictions could be achieved by clustering analysis of a *population* of docking poses. Thus, it is possible to define various bound 'states' (clusters), each comprised of

multiple ‘microstates’ (conformations/poses). The evaluation of the binding energy at each state can then be based both on the energetic interaction *and* on the population of microstates in each state. Indeed the interaction energies, usually accounted for by the scoring functions of the docking software, simply represent the enthalpic contribution to the binding free energy, and the entropic contributions scale with the size of the populations of microstates.

To account for conformational effects, we generate a large numbers (around or more than 1000, which is typically 10-100) of docking poses in both unbiased and biased docking simulations and perform cluster analysis of docking poses. Agglomerative hierarchical clustering scheme is the preferred method, since it does not require the number of clusters as input, which is not known *a priori*, and is controlled by a single parameter. Initially, in this scheme, each docking pose is considered as a cluster. Iteratively, the closest clusters in space are merged, if the distance between them is lower than a threshold. The distance between clusters may be defined to be the average root-mean-square-distance (RMSD) between all pairs of poses (one from each cluster) or as the maximum RMSD between two clusters based on atomic coordinates. We used average distance metric and set the threshold to be 2 Å, which is a commonly used threshold to consider two poses of a small-molecule similar. RMSD calculations in this procedure consider only the heavy atoms of the compounds. In the visual inspection of docking results, we started evaluating from the most populated docking pose clusters. In each cluster, we visualized docking poses in descending docking score order. We used most populated three to six clusters from focused docking simulations to assess or explain the data on the inhibitory action of the ligand.

2.7 SUPPLEMENTARY METHODS

2.7.1 Electrostatic potential calculations using APBS

The visualization of electrostatic potentials generated by a protein provides insights into binding preferences of the protein (Honig & Nicholls, 1995). Continuum and macroscopic solvent models provide alternate means for explicitly modeling solvent molecules. These models provide a solution to the Poisson-Boltzmann (PB) equation (which accounts for the salt effect and the discontinuity of dielectric on the surface of solute molecules in polar solvent) by applying two different dielectric constants. Visualization tools use programs that provide fast numerical solutions of PB equation (Baker et al., 2001; Grant et al., 2001).

In this work, we used visualization of surface electrostatic potentials to understand the selective behavior of MKP inhibitors, by comparing surface properties of related target proteins. We performed our calculation using Adaptive PB Solver (APBS) (Baker et al., 2001). Protein and water dielectric constant values were set to 2 and 78.54, respectively The APBS software requires to submit input files for proteins in PQR format, which contain atomic partial charges and radii in addition to coordinates. These input files were prepared using the web server http://pdb2pqr-1.wustl.edu/pdb2pqr/ (Dolinsky et al., 2004). The surface potential was visualized using VMD (Humphrey et al., 1996). Color scale data range option was set to show from -10 to 10. The color scale data range was set from -10 to 10.

Programs that provide solutions to PB equation are also used in calculating protein-ligand binding free energies (Grant et al., 2001; Brown & Muchmore, 2006). Prospects concerning the use of these methods will be discussed in the last section.

2.7.2 Scientific programming using Python

In much of this work, custom Python programming scripts were developed to perform particular tasks. Python (http://www.python.org) is a popular programming language in computational sciences which is enriched with numerous open-source scientific packages that are readily available to perform computational biology tasks. We used Python and such packages to perform the post-docking clustering task described above, to glue different modeling and simulation tools together by converting the output of one program into an input to another program, or to perform numerical calculations such as PCA and ANM. Some of the tasks and the Python packages that we benefited from are described below.

Exploring sequence and structure databases using Biopython

Biopython (http://biopython.org) is a package containing libraries and functions to perform common bioinformatics tasks. We benefited from Blast module in automated searching of PDB. X-ray structure datasets of drug targets were assembled using these modules.

Numerical calculations using Numpy and Scipy

Scipy and Numpy packages (http://www.scipy.org/SciPy) provide easy access to powerful libraries to manipulate N-dimensional data arrays. We used Numpy and Scipy objects

and functions to manipulate protein coordinate data. Covariance matrices for PCA were calculated using Numpy arrays. Diagonalization of covariance and Hessian matrices were performed using a linear algebra module that uses fast LAPACK routine (Anderson et al., 1999).

Protein structure manipulation and comparison using MMTK

The Molecular Modelling Toolkit (http://dirac.cnrs-orleans.fr/MMTK/) is a package for molecular simulations (Hinsen, 2000). It becomes incredibly convenient to handle protein structures with this package. MMTK was used to read structural ensembles, perform iterative superimposition, and ANM calculation tasks.

Calculation of small-molecule similarity indices using Pybel

The similarity between pairs of small-molecules is calculated by comparing their fingerprints (Daylight, 2007). Molecular fingerprints are represented with binary arrays. Elements of a fingerprint show the absence (0) or existence (1) of chemical features in the molecule. Such chemical features may be existence of an element, a certain atom type, a bond type, or an organic functional group in the molecule. A predefined set of chemical features are used to construct fingerprints for compounds so that each compound has a fingerprint (array) of comparable size.

In this work, the fingerprints and similarities of small-molecule inhibitors found in the X-ray structure datasets were evaluated using the Python implementation of OpenBabel (O'Boyle et al., 2008). fingerprint calculations, we used the FP4 option in Pybel, which uses chemical patterns defining functional groups. The Tanimoto index was used to measure the similarity between two molecular fingerprints, each corresponding to a set of chemical features. The

Tanimoto index is defined as the size of the intersection divided by the size of the union for given feature sets. Hence, this index ranges from 0 to 1, and a value of 1 means a pair of identical fingerprints. When searching molecular libraries, generally a Tanimoto index 0.8 or above is used to retrieve similar compounds. An average number of Tanimoto indices lower than 0.5 for a given set of compounds points to considerable diversity.

3.0 THE INTRINSIC DYNAMICS OF ENZYMES PLAYS A DOMINANT ROLE IN DETERMINING THE STRUCTURAL CHANGES OBSERVED IN INHIBITOR BOUND STRUCTURES

3.1 ANALYSIS OF X-RAY STRUCTURAL ENSEMBLES FOR ENZYMES TARGETED BY DRUGS

We are interested in understanding the determinants of conformational changes observed in drug-bound proteins. To identify a set of targets with multiple structures, we started by examining all proteins listed in two publicly available target databases:

i. The Binding Database (http://www.bindingdb.org/) (Chen et al., 2001) and
ii. DrugBank (http://www.drugbank.ca/) (Wishart et al., 2008).

From these two databases, we obtained a total of more than 5,000 unique target sequences. For each sequence, we searched matching structures in the PDB using automated programming scripts (see subsection 2.7.2). For targets with more than 30 structures, we determined whether

they were crystallized in complexes with different ligands, inhibitors, or substrates. For targets with heterogeneous PDB structure content, we performed automated structural alignment and calculated the RMSD distributions as shown in **Figure 3.1**. Some of the targets that showed considerable backbone variability were HIV-1 reverse transcriptase (RT) (112 structures), p38 mitogen-activated protein (MAP) kinase (p38) (74 structures), cyclin-dependent kinase 2 (Cdk2) (106 structures, excluding 65 Cyclin bound structures), and cAMP dependent protein kinase (81 structures). Targets bound to a heterogeneous set of ligands but displaying limited backbone variability included protein tyrosine phosphatase 1B, lysozyme T4, and carbonic anhydrase.

We considered RT, p38, and Cdk2, which are widely studied as drug targets, in our study. For these three enzymes, we have approximately 300 structures. This large amount of structural data permitted us to perform an extensive comparative analysis of the conformational space being accessed while binding different ligands, and its relation to the intrinsic dynamics of the protein, predicted by the ANM analysis of the unliganded structure. The questions we ask are:

i. How heterogeneous and diverse are the structures of the same protein in different complexes? Here, heterogeneity and diversity refer to protein backbone structural variability and ligand diversity, respectively.

ii. Can we describe the heterogeneity in terms of a few dominant modes accessible under equilibrium conditions, which can be extracted by PCA of the set of structures?

iii. To what extent does the protein select from this pre-existing equilibrium when binding the ligand; or to what extent does the particular ligand induce a substrate-specific rearrangement?

First, we discuss structural variability and small-molecule diversity in the datasets using two common and simple measures. Then, we present PCA and ANM results for each of these target proteins.

3.1.1 Structural Ensembles of drug target enzymes

On the structural variability in the datasets

To display the conformational variability in these datasets, we show RMSD distributions in **Figure 3.1**. The PDB IDs of structures in these ensembles are listed in the **Appendix A** tables **Table A.1**, **Table A.2**, and **Table A.3**. Left panels in **Figure 3.1** show the RMSD distributions with respect to a reference structure. In each case, the reference structure is chosen to be an unliganded form of the protein. PDB IDs of reference structures are given in respective **Appendix A** tables. Right panels show all-to-all RMSD distributions. Structures in the RT ensemble display largest deviations, which is primarily due to the large size of this protein (998 residues) and the high mobility of its domains. The p38 dataset showed moderate flexibility (354 residues), and the Cdk2 (298 residues) dataset showed the lowest variability.

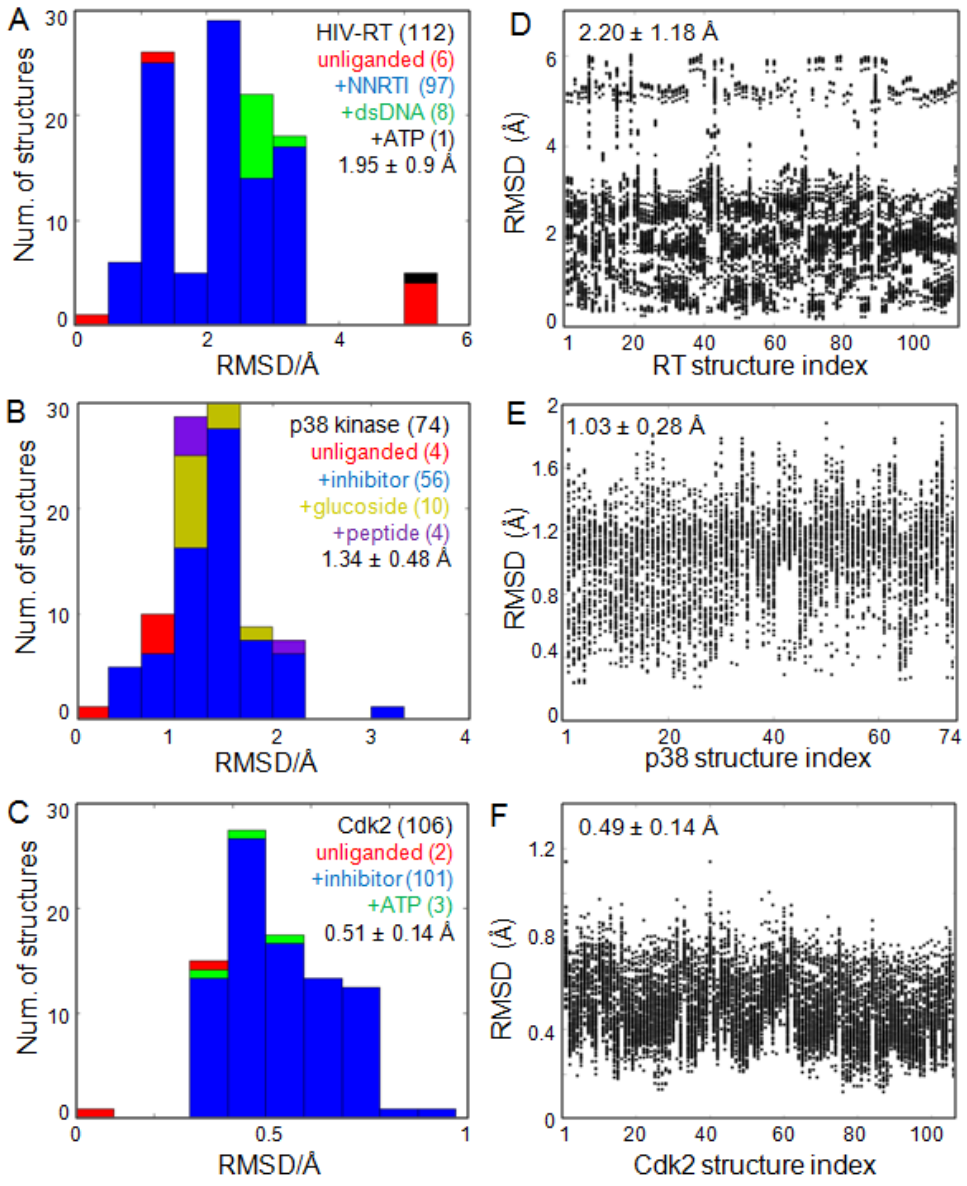


Figure 3.1 RMSD distributions for three sets of structures: HIV-1 RT, p38 and Cdk2.

Panels A, B, and C show the RMSD distributions with respect to the respective reference structures for datasets I, II, and III. The Cα RMSDs are based on the residues common to the reference structure and each superimposed structure. Panels D, E and F show the RMSD values for all pairs of structures in the dataset, for each member of the set indexed (alphabetically ordered according to their PDB identifiers) along the abscissa. Average pairwise RMSD values and standard deviations are given in each panel. Figure is adopted from (Bakan & Bahar, 2009).

On the diversity of bound inhibitors in the datasets

Also notable is the structural diversity of the bound small-molecules in each dataset. To show this, we calculated the pairwise similarities by averaging the Tanimoto indices over all pairs of compounds (**Figure 3.2**; see also subsection 2.7.2). The RT dataset contained total of 59 different NNRTIs with an average similarity index of 0.37 ± 0.13. The p38 dataset contained 55 compounds with an average index of 0.40 ± 0.15. The Cdk2 dataset contained 100 different compounds with an average index of 0.43 ± 0.14. These values show that RT, p38, and Cdk2 datasets contained considerably diverse ensembles of inhibitors. As will be shown in the following subsections, the diversity of bound inhibitors are presumed to effect conformational variability observed in these datasets for one of these two reasons: (i) the shape of the binding sites is directly affected by the dominant modes of variability in the dataset (p38 and Cdk2 datasets), and (ii) the binding site is located at a hinge site that controls the collective dynamics of the target protein (RT dataset). Yet to be answered is whether the binding of the small-molecule induces these changes or it is the protein that samples those conformations even in the absence of the bound small molecules. Our analysis will clarify this issue.

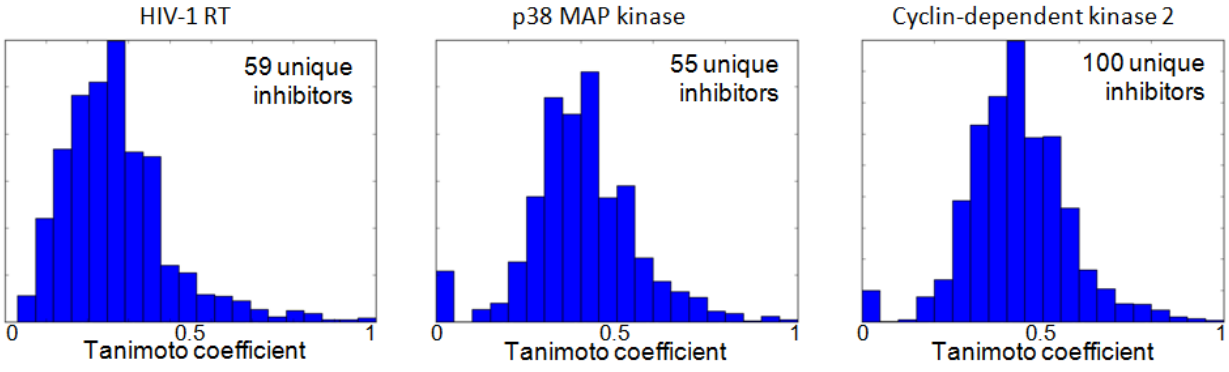


Figure 3.2 Distribution of pair-wise Tanimoto coefficients for inhibitors in X-ray datasets.

3.1.2 HIV-1 reverse transcriptase

HIV-1 RT is a multifunctional enzyme composed of two subunits, p66 and p51 (Kohlstaedt et al., 1992). p66 contains the polymerase and RNase H domains. The polymerase domain is described as a right-hand containing fingers, palm, thumb, and connection subdomains. The connection subdomain links the polymerase and RNase H domains. The p51 subunit bears only the polymerase domain, comprised of the same subdomains, arranged in a different tertiary structure. Non-nucleoside RT inhibitors (NNRTIs) are known to act allosterically by interfering with the functional motions of the p66 fingers and thumb upon binding a pocket at a global hinge site on the palm subdomain (Kohlstaedt et al., 1992; Bahar et al., 1999; Temiz & Bahar, 2002; Zhou et al., 2005).

PCA results

Figure 3.3A shows the projection of RT structures onto the subspace spanned by the first two principal axes, PC1 and PC2, determined for the examined dataset (**Table A.1**). The points therein represent 112 RT structures (see subsection 2.1.3). These two PCA modes were found to account for 71% of the total variance in structure (Equation 2.1.7). Notably PC1 provides a clear separation of the structures into three clusters according to the types of ligands. ATP bound structure is found among the unliganded structures, since ATP binding doesn't have an affect on global shape of the protein. Two unliganded structures are also found among the NNRTI bound structures. These structures were obtained after bound NNRTI is soaked out [REF], and shows that unliganded RT remains stable in NNRTI bound conformational space.

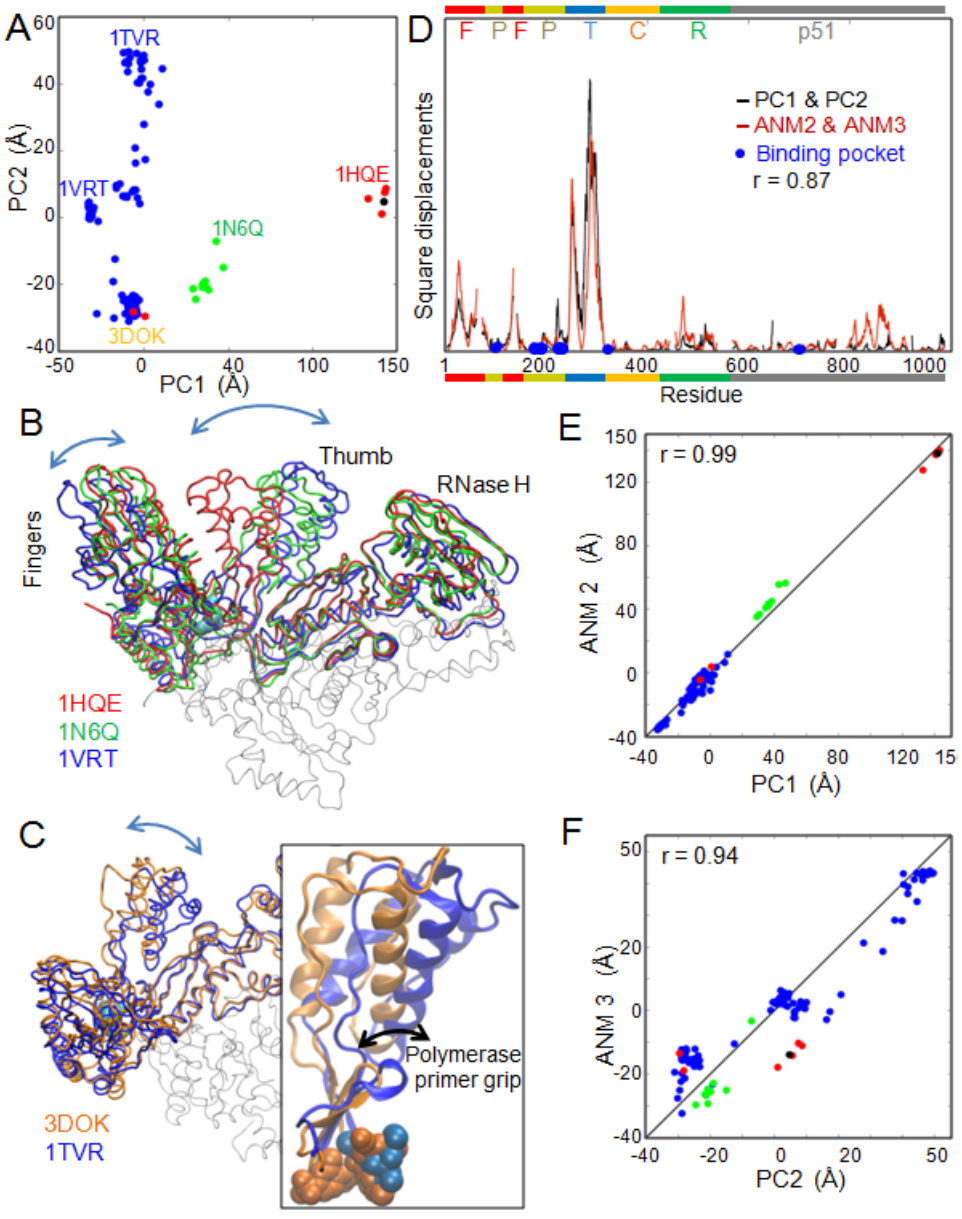


Figure 3.3 PCA and ANM results for HIV-1 RT.

(A) Projection of 6 unliganded (red), 97 NNRTI-bound (blue), 8 dsDNA/RNA-bound (green), and 1 ATP-bound (black) RT structures onto PC1 and PC2. (B) Structural variation along PC1, illustrated using selected structures labeled in panel A. (C) Structural variation along PC2. The inset shows a closer view of the thumb subdomain. Inhibitors are shown by the same color as the corresponding RT conformation. In both panels B and C, thumb domain moves relatively most, but along orthogonal directions. (D) Comparison of the weighted sum of square displacements along PC1 and PC2, with those along ANM modes 2 and 3 predicted for unliganded reference structure. (E) Projections of the 112 structures onto PC1 and ANM2 directions. Standard deviation along this axis is

37.2 Å. (F) Projections onto PC2 and ANM3. Standard deviation along this axis is 26.4 Å. Figure is adopted from (Bakan & Bahar, 2009).

The structural variation represented by PC1 is shown in **Figure 3.3B**. The most distinctive feature is the large movement of the thumb. More careful examination reveals the anti-correlated displacements of the fingers and thumb of about 10Å and 20Å at their most exposed regions, respectively. The fluctuations along this mode contribute by 47% to the total variance, and span a cumulative displacement of 176Å along this axis (this number is the difference between coordinates of structures that fall on the extremes of this principal axis). Note that this value refers to the cumulative displacement summed over all residues. This is a strikingly large number, hence is illustrated in **Figure 3.4** (see also subsection 2.1.3 for a description). The tendency of RT to sample conformations along this mode in the absence of ligands is evidenced by site-directed spin labeling experiments (Kensch et al., 2000) and supported by ANM (Temiz & Bahar, 2002) and MD simulations (Ivetac & McCammon, 2009).

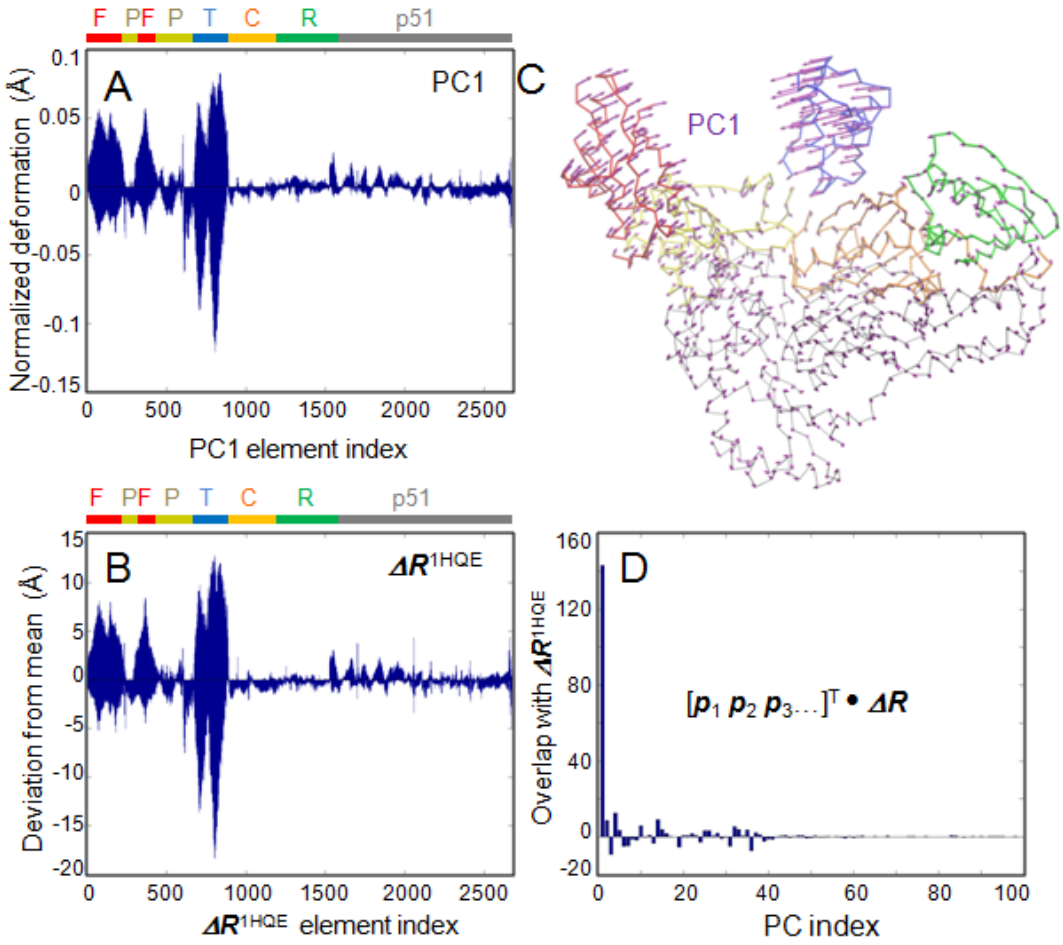


Figure 3.4 Projection of conformational changes onto PC1.

(A) Plot of PC1 (components of the *3N*-dimensional vector $\boldsymbol{p}_k$) obtained from the PCA of the RT dataset. (B) Plot of the *3N*-dimensional conformational change/deformation vector, ΔR^{1HQE} exhibited by 1HQE with respect to the mean structure in the ensemble. Note the close similarity between the shapes of $\boldsymbol{p}_k$ and ΔR^{1HQE} in the respective panels A and B, indicating that the structural difference of 1HQE with respect to the mean structure is almost fully explained by a collective displacement along PC1 (see panel D). This is because PC1 captures the largest magnitude differences, and for this dataset unliganded structures show the largest deviation from the mean structure. The magnitude $\|\Delta R^{1HQE}\|$ of the conformational change vector for 1QHE is 145.9 Å. The RMSD of this structure from the mean coordinates is 4.89 Å. (C) Directions of motions along PC1 shown on the average set of coordinates. (D) Contribution of different PCA modes to the deformation of 1HQE with respect to the mean structure, represented by the projection c_i^{1HQE} of ΔR^{1HQE} onto $\boldsymbol{p}_k$. Note that the projection of PC1, c_1^{1HQE} is considerably large in this case

(c_1^{1HQE} = 143.3 Å) as the deformation of the unliganded structure 1HQE is almost fully along the principal direction $\boldsymbol{p}_k$. Figure is adopted from (Bakan & Bahar, 2009).

The second principal mode, PC2, predicted by the PCA of RT structures describes the out-of-plane fluctuations of the thumb (**Figure 3.3C**), which would not be obvious from comparisons of arbitrarily chosen structures. These fluctuations are orthogonal to those described by PC1. Resulting structural differences are illustrated in **Figure 3.3C** using two structures separated by 81 Å. The inset shows a close up view of the thumb. The thumb and the polymerase primer grip (part of the palm) move together as a rigid body. Motions of thumb domain and residues at the tip of polymerase grip are important for polymerase activity. Impairing their functional motions and positioning is how NNRTI inhibitors act.

As a further test, we performed the PCA of the NNRTI-bound subset. The PC1 in this case was found to be almost identical (correlation coefficient of 0.99) to the PC2 of the complete set, and contributed by 50% to the total variance (**Table 3.1**). It was also interesting to note that the PC1 from the complete ensemble did not have a strong counterpart in the NNRTI-bound subset. This supports the view that NNRTI binding depresses the anti-correlated fluctuations of the fingers and thumb, and stimulates thumb fluctuations in an orthogonal direction. This observation is in parallel with the previously proposed view that NNRTI inhibition is achieved by imparting a change in the direction of the thumb movements (Temiz & Bahar, 2002).

Table 3.1 Overlap between PCA modes obtained from complete ensembles of reverse transcriptase (RT) structures with those obtained from NNRTI bound subset of the ensemble

PCA modes (fractional contribution)		Subset of NNRTI-bound structures		
		PC1 (0.50)	PC2 (0.18)	PC3 (0.08)
All RT structures	PC1 (0.47)	0.07	0.58	0.50
	PC2 (0.24)	0.99	0.00	0.09
	PC3 (0.09)	0.11	0.60	0.54

How does a small molecule perturb the global dynamics of such a large structure? The answer lies in the location of the NNRTI binding pocket. As shown in the displacement profile in **Figure 3.3D**, the binding pocket residues show minimal, if any, variations in their positions in these two dominant PCA modes. ANM calculations presented below also confirm that the NNRTI binding residues are severely constrained in the global modes of RT. Perturbation of such constrained regions in the global modes may have a dramatic effect on functional dynamics. Yet we not that inhibition of function by blocking a global motion require a druggable binding site. The special location of NNRTI binding site and its druggable physicochemical properties enables NNRTIs work effectively.

Table 3.2 Overlap between PCA and ANM modes for the ensemble of RT structures

PCA modes		ANM1	ANM2	ANM3
HIV-RT	PC1 (0.47)	0.25	0.89	0.11
	PC2 (0.24)	0.47	0.08	0.63
	PC3 (0.09)	0.42	0.06	0.15

Comparison with ANM results

Figure 3.3D displays the joint effect of ANM modes 2 and 3. These two ANM modes were found to yield the highest correlation (among all ANM modes) with PC1 and PC2, respectively (see **Table 3.2**). ANM mode 1, on the other hand, refers to the anticorrelated

fluctuations of fingers subdomain and RNase H domain. Comparison with PCA modes showed that this mode shows a weak correlation (0.52) with the PC5. The directions of the first three ANM modes are shown in **Figure 3.5**.

As a direct comparison of the structural changes along PC1 and the motions induced in ANM mode 2 (ANM2), we examined the level of correlation between the projections of the structures onto these two collective displacement directions. **Figure 3.3E** displays the results. Strikingly, the structures perfectly align along these two axes (correlation coefficient of 0.99), demonstrating the equivalence of the predicted (ANM2) and experimentally observed (PC1) global modes. Similarly, by projecting the structures onto ANM3, and its PCA counterpart, PC2, we find a correlation coefficient of 0.94, again supporting the view that the most distinctive structural changes assumed by the NNRTI-bound RTs simply originate from intrinsically favorable ANM modes (**Figure 3.3F**).

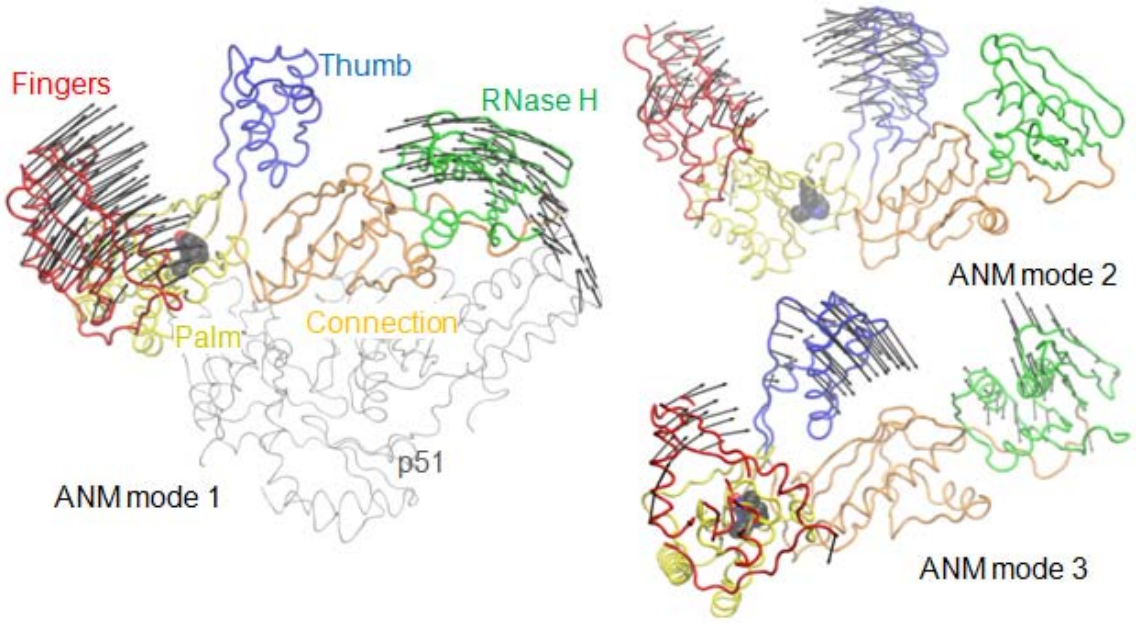


Figure 3.5 Top three ANM modes for reference RT structure.

The structure determined by Esnouf et al. (Esnouf et al., 1995) (PDB id: 1RTJ) was used in calculations. The left diagram displays the intact RT, comprised of subunits p66 and p51. The p51 subunit is not shown in the right diagrams as it is almost rigid in those modes. The p66 subunit domains/subdomains (thumb, finger, palm and connection subdomains on the polymerase domain, and the RNase H domain) are shown in different colors (blue, red, yellow, orange/brown, green and gray, respectively) for visual clarity. The arrows indicate the directions and relative sizes of motions predicted in the ANM modes 1 (top), 2 (middle) and 3 (bottom). Figure is adapted from (Bakan & Bahar, 2009).

When put together, these results suggest that RT samples conformations pre-disposed to NNRTI binding, and NNRTI binding shifts RT dynamics from one mode to another, both being intrinsically favored.

3.1.3 p38 MAP kinase

The p38s MAP kinases are serine/threonine kinases activated in response to external stress. They regulate the production of proinflammatory cytokines, and hence serve as targets in the treatment of inflammatory diseases (Kumar et al., 2003). The structure and interactions of p38s with different inhibitor classes are well characterized. Their dynamics, however, is not well understood and poses challenges in inhibitor design (Cavasotto & Abagyan, 2004).

PCA results

The results from PCA are presented in **Figure 3.6**. Panel **A** displays the distribution of 74 structures of p38 isoform α. The p38 structure has a canonical kinase fold composed of a β-

sheets rich N-terminal lobe (N-lobe) and an α-helix rich C-terminal lobe (C-lobe) (Wang et al., 1997). The catalytic site, and also the binding site for competitive inhibitors, is the cleft between these two lobes. Example structures from different subgroups are displayed in panels B and C to illustrate the major structural changes along PC1 and PC2, respectively.

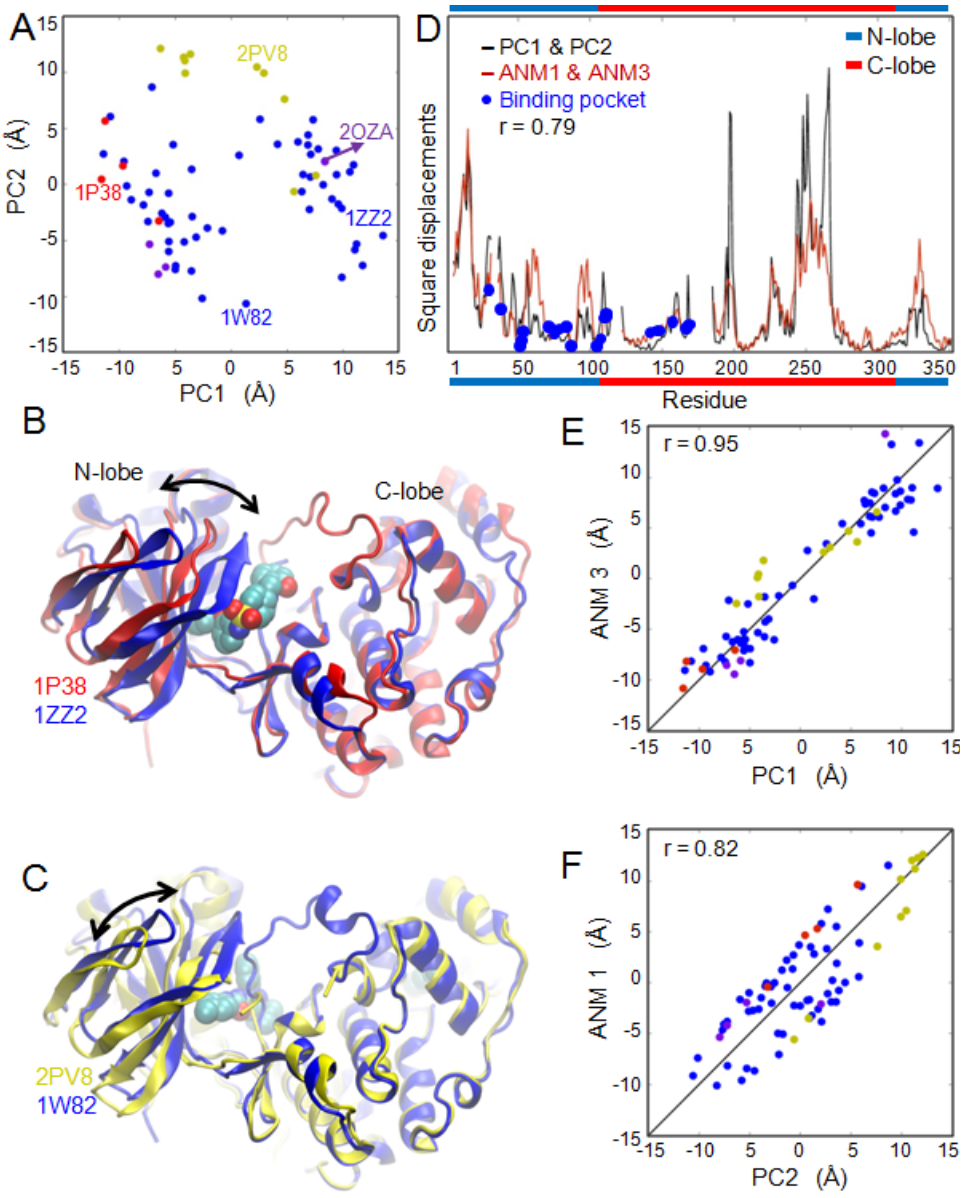


Figure 3.6 PCA and ANM results for p38 MAP kinase.

(A) Projection of 4 unliganded (red dots), and 56 inhibitor-bound (blue), 10 glucoside-bound (yellow), and 4 peptide-bound (violet) p38 structures onto PC1 and PC2. Distant structures along PC1 and PC2 are selected to illustrate in the respective panels (B) and (C) the structural variations represented by these PCs. Note that only C-lobes are considered in structural alignment for visualization purposes. (D) Square displacements of residues along PC1 and PC2, compared to those driven by ANM modes 1 and 3. (E) Projections of the 74 structures onto PC1 and ANM3. (F) Projections onto PC2 and ANM1. Figure is adopted from (Bakan & Bahar, 2009).

PC1 refers to anti-correlated movements of the two lobes, as shown by the superimposition of an unliganded (red) and an inhibitor-bound (blue) structure in panel B. These movements map to a separation of more than 25 Å between the two conformers along PC1 axis (panel A). PC2, on the other hand, involves twisting motion of N-lobe with respect to the C-lobe illustrated by a glucoside- and an inhibitor-bound structure (panel C). The size of the motions along PC2 is comparable to that along PC1 (panel A). The global backbone (Cα) changes observed in the ensemble are well described by the first two PCs. Notably, although these modes represent only 2 degrees of freedom out of a total of 963, they account for 45% of the variance in the backbone structure observed in the dataset (**Table 3.3**), and reconfigurations along these two directions allow for reducing the average RMSD with respect to the mean structure from 1.0 ± 0.3 Å to 0.53 ± 0.17 Å.

Table 3.3 Overlap between PCA and ANM modes for the ensemble of p38 structures

PCA modes		ANM1	ANM2	ANM3
p38	PC1 (0.29)	0.39	0.05	0.71
	PC2 (0.16)	0.79	0.21	0.22
	PC3 (0.11)	0.06	0.57	0.05

Fluctuations along PC1 determine the exposure of the binding cleft. In the unliganded state, the cleft is wide open, presumably to facilitate ATP/inhibitor recognition and optimal interaction. Upon ligand binding the cleft closes down. The closure of the cleft increases the packing interactions with the bound molecule. In one of the peptide-bound structures, p38 is complexed with its primary substrate MAPK-activated protein kinase 2 (MK2), which is located 20Å away from the unliganded p38 along PC1 (labeled as 2OZA in **Figure 3.6A**). Other peptide-bound structures contain short stretches, corresponding to docking sites of p38 regulators/substrates, with comparably much less impact on p38. The conformational change observed when p38 is crystallized with MK2 potentially affects its function (White et al., 2007). These show that movements captured by PC1 are functionally relevant. Fluctuations along PC2, on the other hand, change the relative positioning of the atoms lining the cleft between the lobes, affecting in particular the β-strands 1 to 3 and the connecting hairpins.

Figure 3.6D displays the square displacements of residues resulting from the weighted contributions of movements along PC1 and PC2. The binding pocket residues are marked by the blue dots. Some of them, corresponding to hinge sites in these global modes, are extremely constrained, and others show moderate displacement. The structural diversity of inhibitors may be a primary reason for the conformational heterogeneity along these modes. The average pairwise similarity among the 55 compounds in the dataset is 0.40 ± 0.15 based on standard cheminformatics metrics. Apparently, the movements of the lobes allow for optimizing their interactions with the small-molecules, and hence the poor results in *in silico* cross-docking experiments, when the target backbone is assumed to be rigid (Cavasotto & Abagyan, 2004).

Comparison with ANM results

The correlations between the lowest three ANM modes and the top-ranking three PC directions are listed in **Table 3.3**. The counterparts of PC1 and PC2 are found to be ANM3 and ANM1, with respective correlation coefficients of 0.71 and 0.79. ANM2, a twisting motion of the two lobes, was found to correspond to PC3 with a coefficient 0.57 (**Figure 3.7** illustrates these three ANM modes). The displacements of residues induced by ANM modes 1 and 3 are compared in **Figure 3.6D** with those driven by PC1 and PC2. The two profiles agree with a correlation coefficient of 0.79. Also important is the directionality of ANM-predicted fluctuations. Hence, we projected the ensemble onto these ANM modes and distributions along corresponding PCs. Distributions of the structures along the PC1-ANM3 (**Figure 3.6E**) and PC2-ANM1 (**Figure 3.6F**) yielded correlation coefficients of 0.95 and 0.82, respectively. This excellent correspondence underscores the robustness of the low-frequency modes, and shows their functional importance in accommodating the binding of structurally diverse inhibitors as well as the cellular protein substrate MK2.

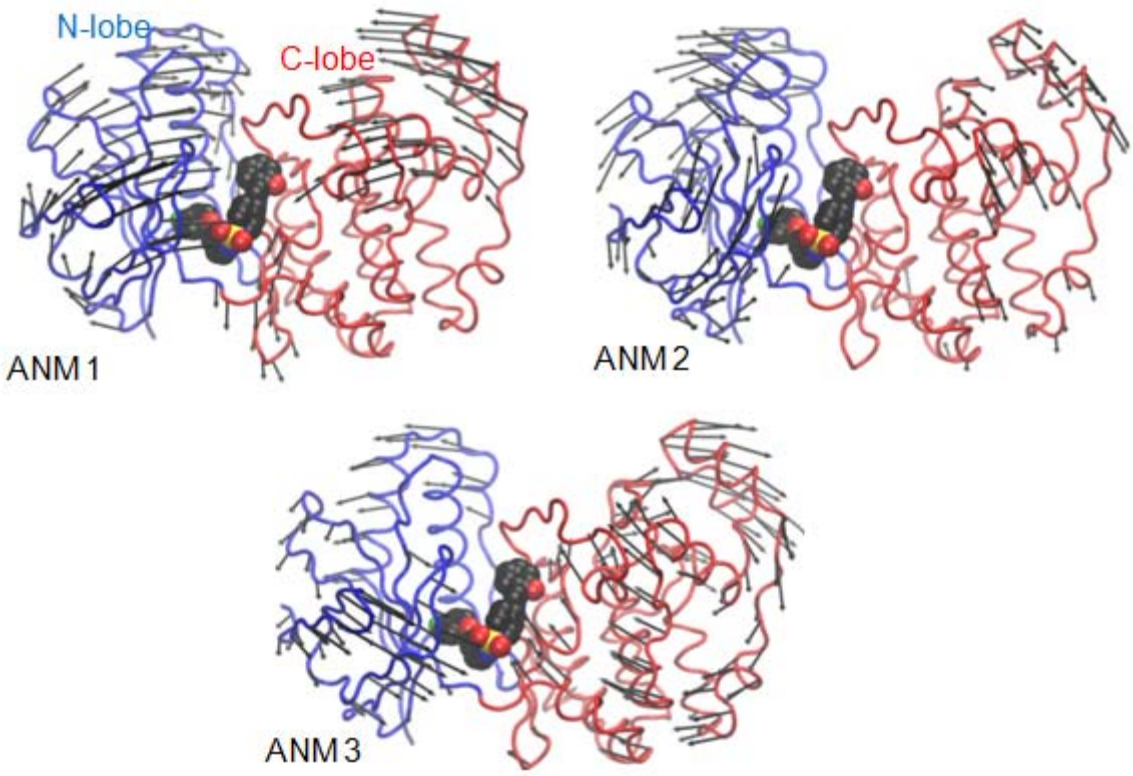


Figure 3.7 Top three ANM modes for reference p38 structure.

Calculations were performed using the structure determined by Wang et al. (Wang et al., 1997) (PDB id: 1P38). The central small-molecule, shown in space-filling, was not included in the calculations, but displayed in the figure to illustrate the ligand binding site. Figure is adapted from (Bakan & Bahar, 2009).

3.1.4 Cyclin-dependent kinase 2

Cdks are serine/threonine kinases involved in the regulation and progression of the cell cycle. Cdk activity is regulated by activator (cyclin family) or inhibitor (INK and Cip families) protein binding and phosphorylation (Pavletich, 1999). In cyclin- or Ink-bound structures, the N-lobe adopts a host of distinct conformations intrinsically favored by Cdk (Tobi & Bahar, 2005). This

conformational flexibility impacts the ATP/small-molecule pocket, and hence needs to be considered in designing drugs (Cavasotto & Abagyan, 2004).

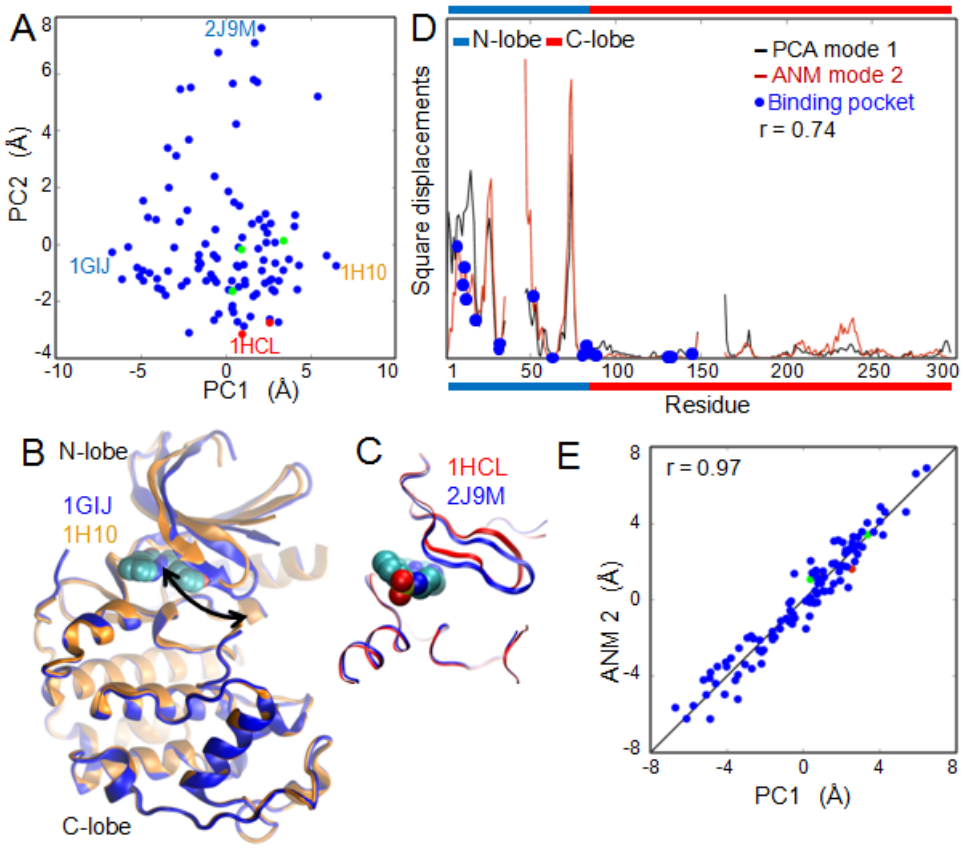


Figure 3.8 PCA and ANM results for Cdk2.

(A) Projection of 2 unliganded (red), 3 ATP-bound (green), and 101 inhibitor-bound (blue) Cdk2 structures onto PC1 and PC2. (B) Structural variation along PC1. Note that only C-lobes are considered in structural alignment for visualization purposes. (C) Structural variation along PC2. (D) Comparison of the square displacements of residues along PC1 and ANM2. (E) Projection of 106 Cdk2 structures onto PC1 and ANM2. Figure is adopted from (Bakan & Bahar, 2009).

Results from PCA

We have assembled a set of 106 Cdk2 structures. This set was more narrowly distributed compared to the previous two cases (average RMSD of 0.50 ± 0.14 Å with respect to the reference structure; **Figure 3.1**) and therefore the structural variations may not be large and precise enough to classify them into distinctive PC modes. Yet, the first two PCA modes were able to account for 39% of the variance in the dataset (**Table 3.4**). The projection of Cdk2 structures onto PC1 and PC2 showed a diffuse distribution, not clearly distinguishing the different bound forms, although the two unliganded structures clustered together at a distinctive end of the subspace (**Figure 3.8**). We selected the Cdk2 structures that fall in the extremes of this distribution to visualize the conformational differences along these two PCA directions. PC1 describes the twisting motion of N- and C-lobes, which is comparable to p38 movements along PC2. This is illustrated by two structures that are 13.2 Å apart along this axis (**Figure 3.8B**). In **Figure 3.8A**, the unliganded structures fall close to the center on PC1 coordinate, which indicates that twisting motion occurs in either direction. The structural variation described by the PC2, on the other hand, was localized at chain termini and flexible loops. Panel C illustrates such a local movement for the so-called glycine loop that lines the binding cleft. The two structures in this panel are 10.8 Å apart when projected onto PC2. The heterogeneity of the ensemble along these modes presumably originates from the physicochemical diversity of bound inhibitors. The ensemble contained 100 different compounds with an average pairwise similarity metric of 0.43 ± 0.14.

Table 3.4 Overlap between PCA and ANM modes for the ensemble of Cdk2 structures

PCA modes		ANM1	ANM2	ANM3
Cdk2	PC1 (0.23)	0.36	0.73	0.09
	PC2 (0.16)	0.06	0.05	0.26
	PC3 (0.12)	0.48	0.09	0.34

Comparison with ANM predictions

The top ranking ANM modes predicted for the unliganded structure (1HCL) are displayed in **Figure 3.9**. Among them, ANM2 yields a correlation of 0.73 with PC1 (**Table 3.4**). The square displacements of residues along these modes are compared in **Figure 3.8D**. Some of the binding pocket residues are located at the minima of these profiles, while the rest are located in highly mobile regions of N-lobe. Those at the minima include Val64, Phe80, Asn132 and Ala144-Asp145, which are buried deep into the cleft between the two lobes, while those exhibiting moderate-to-high flexibility are at the periphery of the cleft. This indicates that fluctuations along this mode allow for the positioning of recognition site residues, while other functional residues, such as Asn132 and Asp145 that coordinate the metal atom for the catalytic reaction, and the catalytic residues Asp127 and Lys129 are almost fixed. The projections of the ensemble of structures onto ANM2 and PC1 yield a correlation of 0.97 (panel E). These suggest that despite the minimal changes in structure, the observed structural heterogeneity is not random, but geared towards the intrinsically favored mode of motion. The changes along the experimentally observed PC2, however, seem to be rather local and would rather be viewed as induced by inhibitor binding.

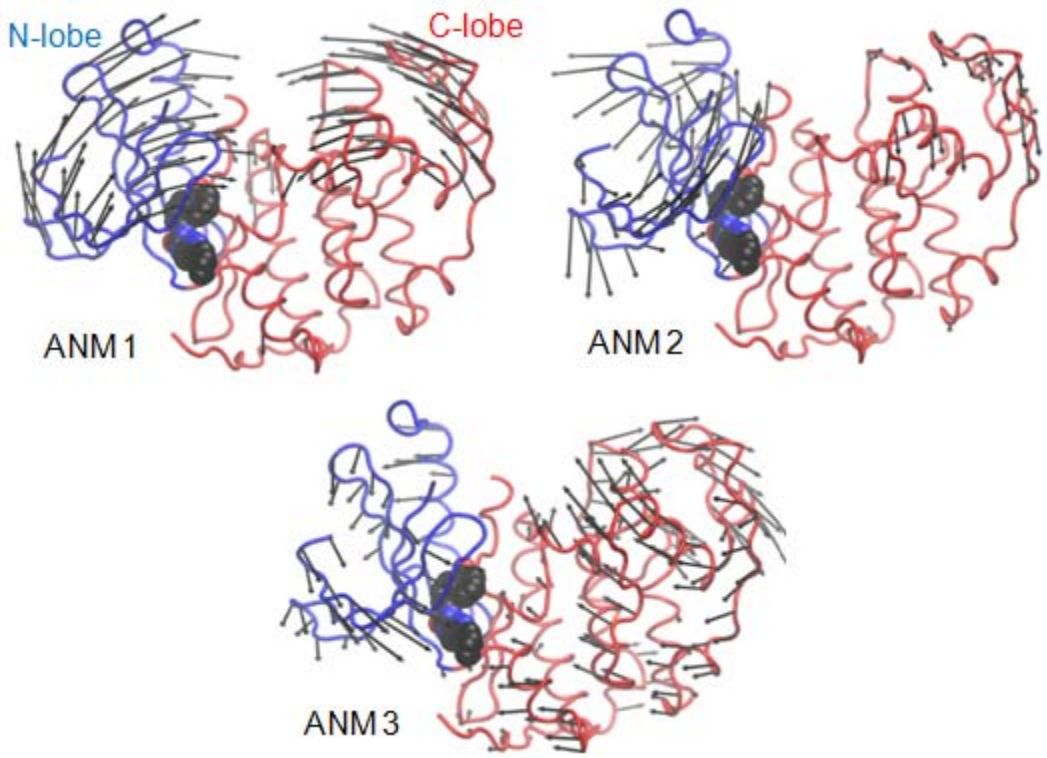


Figure 3.9 Top three ANM modes for reference Cdk2 structure.

ANM calculations were performed for the structure resolved by Schulze-Gahmen et al. (Schulze-Gahmen et al., 1995) (PDB id: 1HCL). A small molecule (not included in the calculations) is shown in space filling representation to indicate the ligand binding site. Figure is adapted from (Bakan & Bahar, 2009).

3.2 SOLUTION DYNAMICS OF PROTEINS COMPARED WITH ANM

So far, we showed that the protein topology offers a few well-defined, energetically favorable, mechanisms/modes of structural rearrangements along preferential mode axes; and the ligand selects the one that best matches its structural and dynamic properties. Yet, this may mean that the ligand may be able to induce observed collective conformal changes, since they are along low-energy mode directions. If we are able to show that these low-energy modes do also describe

the solution dynamics of proteins in the absence of ligands, this will unambiguously demonstrate that the bound small-molecule does *not* induce the observed conformational changes, but selects from a set of conformations that are sampled by the protein even in the absence of the ligand. To explore this, we analyzed broad ensembles of NMR models determined for ubiquitin (Lange et al., 2008) and calmodulin (CaM) (Gsponer et al., 2008). In **Figure 3.10**, we display the RMSD distributions for these two ensembles. Our analysis confirmed that the conformational space sampled by these proteins in solution, in the absence of ligand, is consistent with ANM-predicted global modes, as will be presented in details in this section.

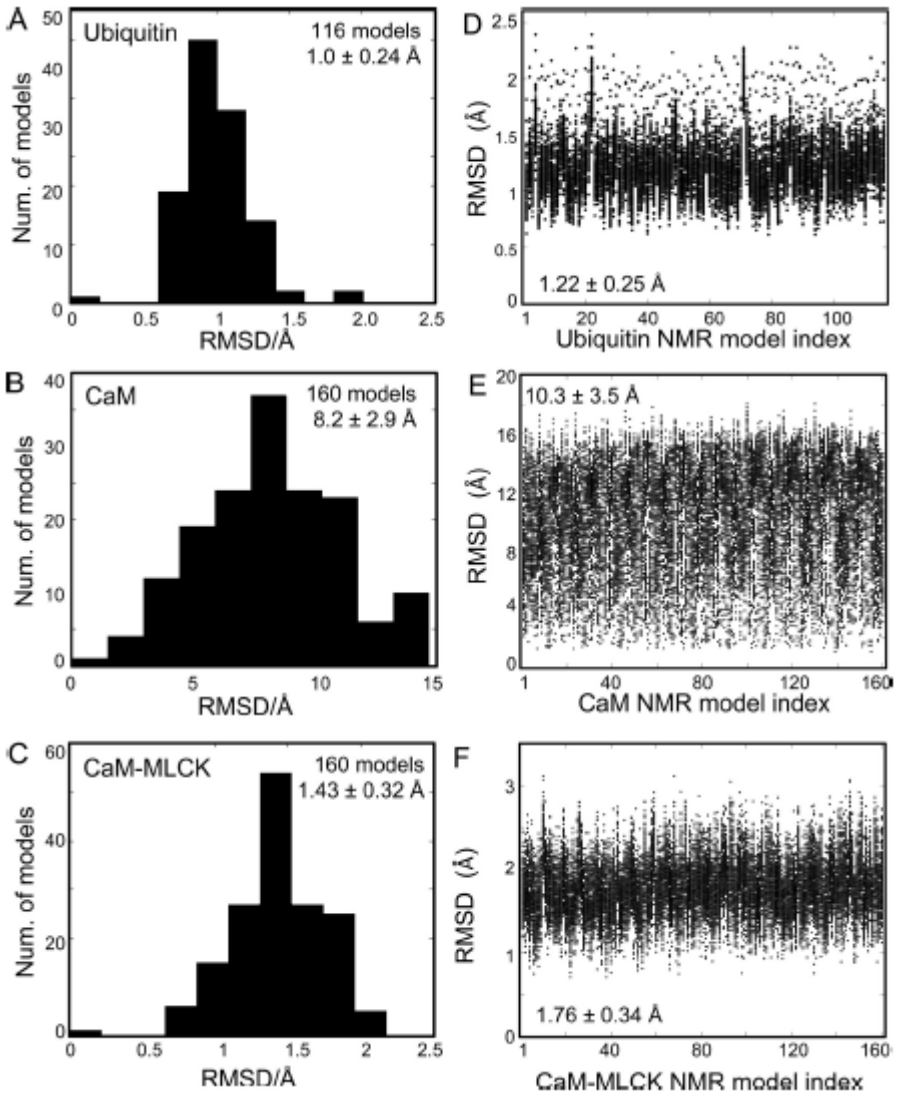


Figure 3.10 RMSD distributions for three ensembles of NMR models.

Ensembles belong to ubiquitin, calmodulin, and calmodulin complexed with a 19-residue peptide from myosin light chain kinase (MLCK). Panels A, B and C show the RMSD distributions with respect to the reference structures in each set (respective PDB codes: 2K39, 2K0E and 2K0F). The number of NMR models in each ensemble is indicated in the panels. Panels D, E and F show the pairwise RMSD values with respect to all other models in the ensemble for each model (indexed along the ordinate), similar to panels D-F of Figure 3.1. The model indices along the abscissa correspond to those assigned in the respective PDB files. Average RMSD values and standard deviations are given in each panel. The results for ubiquitin refer to residues 1-70, excluding the disordered 6 residues at the C-terminus. Note the broad distribution in RMSDs for CaM (panels B and E). Figure is adopted from (Bakan & Bahar, 2009).

The ensembles of NMR models deposited for ubiquitin (Lange et al., 2008) and for CaM (Gsponer et al., 2008) have been pointed out to sample conformations comparable to those observed in their substrate-bound forms. The question is: how do principal modes of structural change extracted from the PCA of these ensembles correlate with those predicted by the ANM?

Results are presented in **Figure 3.11**. The projection of the NMR models onto PC1 and PC2 are shown in both cases to exhibit a close agreement with one or two ANM global modes. We note that CaM samples a very broad conformational space in its unliganded form (**Figure 3.13**) as the N- and C-terminal domains (NTD and CTD) are connected by a helix that is readily flexed. We examined the principal modes accessible to the NTD and CTD in relation to their ANM counterparts, as a further comparison. Correlations in the range 87% to 94% were observed in both cases (**Figure 3.12**, panels A-D), demonstrating that the ANM provides a good representation of the variability of the structures even within domains, provided that they are sufficiently decoupled.

Finally, we also analyzed the complex CaM-MLCK with a myosin light chain kinase (MLCK) peptide, which further confirmed that principal modes derived from NMR models are in accord with ANM predictions (correlations of 0.88 and 0.77 in **Figure 3.12** panels E-F). The results for all examined NMR ensembles are compiled in **Table 3.5** and **Table 3.6**, and illustrated in **Figure 3.13**. The cumulative overlaps between subsets of 3, 6, and 20 ANM modes with the top three PCs (**Table 3.6**) further support the consistent correlation between the

subspace of conformations seen in the experimental structures and those predicted by computations.

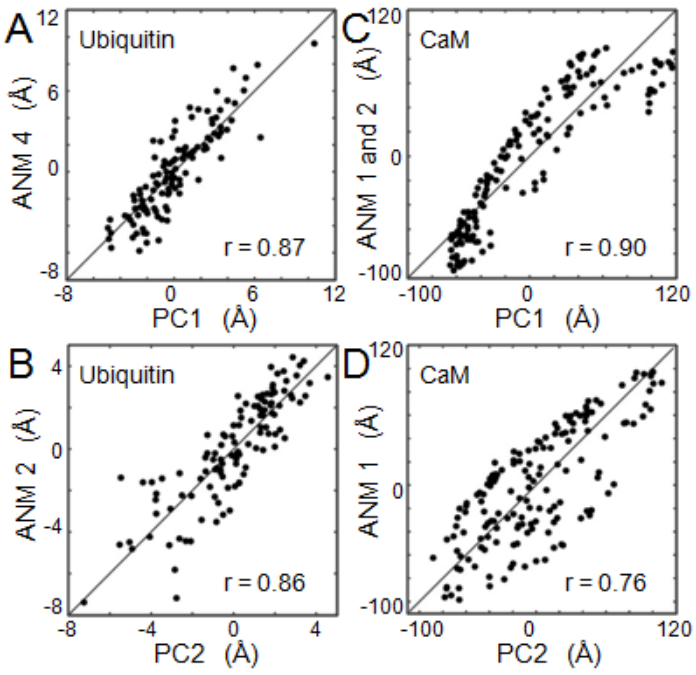


Figure 3.11 Comparison of the PC and ANM modes for ubiquitin and CaM ensembles.

The projections of NMR models for ubiquitin (panels A and B), and CaM (C and D) onto PC1 and PC2 are compared with the projections onto ANM global modes. ANM calculations are performed for the model that has the lowest RMSD with respect to all others in each ensemble. Figure is adopted from (Bakan & Bahar, 2009).

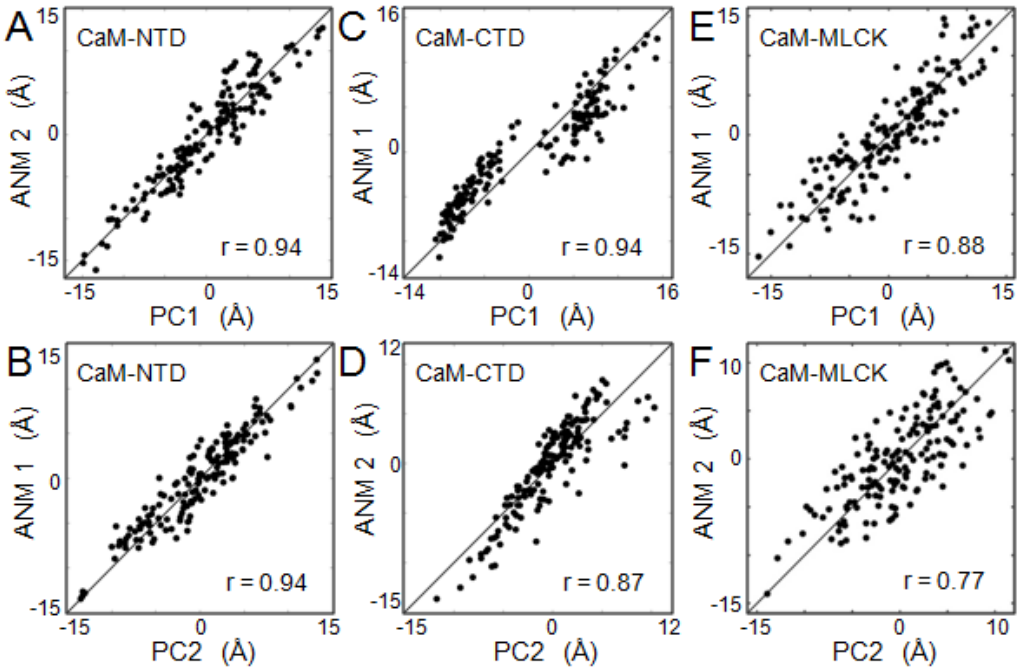


Figure 3.12 Comparison of the PC and ANM modes for CaM ensembles.

Projections of NMR ensemble onto PCA modes 1 and 2 are compared with projections onto ANM modes 1 and 2, for CaM-NTD (panels A and B), CaM-CTD (panels C and D), and CaM-MLCK (panels E and F). Residues 5-79 and residues 83-147 were considered for the respective NTD and CTD, respectively. ANM calculations were performed for a representative model (the one with the lowest RMSD from all others) in each ensemble. The corresponding models numbers are 101, 1, and 21 for CaM-NTD, CaM-CTD, and CaM-CLMK, respectively. Figure is adopted from (Bakan & Bahar, 2009).

Table 3.5 Overlap between PCA and ANM modes for NMR ensembles

PCA modes		ANM1	ANM2	ANM3
Ubiquitin	PC1 (0.13)	0.19	0.30	0.62*
	PC2 (0.09)	0.09	0.72	0.16*
	PC3 (0.08)	0.31	0.06	0.00*
CaM	PC1 (0.31)	0.54	0.54	0.30
	PC2 (0.25)	0.71	0.49	0.17
	PC3 (0.20)	0.05	0.54	0.57
CaM-MLCK	PC1 (0.18)	0.72	0.37	0.12
	PC2 (0.10)	0.21	0.75	0.15
	PC3 (0.09)	0.37	0.28	0.43
CaM-NTD	PC1 (0.25)	0.29	0.70	0.48
	PC2 (0.21)	0.91	0.18	0.23
	PC3 (0.12)	0.05	0.03	0.09
CaM-CTD	PC1 (0.51)	0.61	0.13	0.17
	PC2 (0.14)	0.32	0.70	0.02
	PC3 (0.08)	0.16	0.28	0.29
* Overlap values for ANM mode 4.				

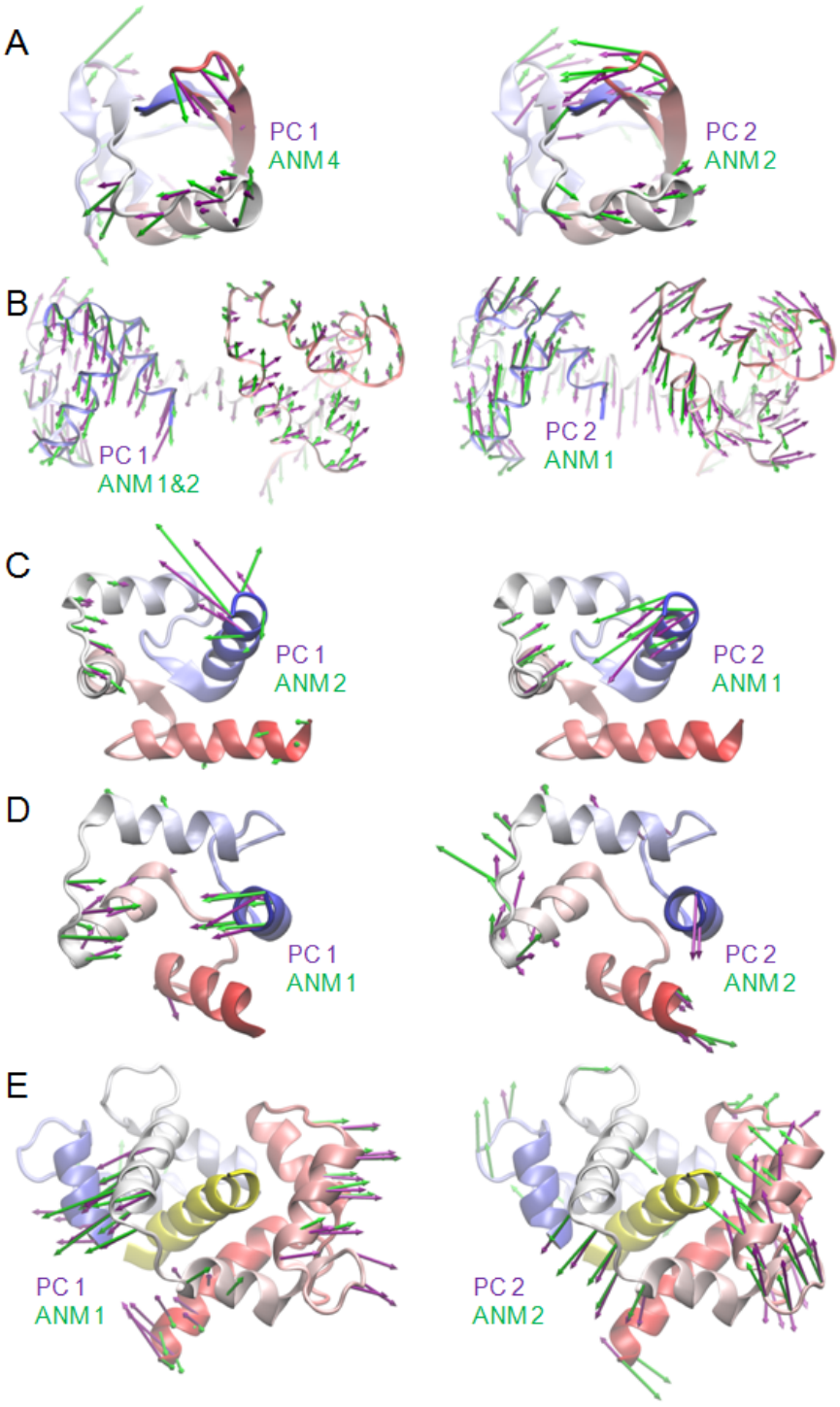


Figure 3.13 Directions of PCA modes and corresponding ANM modes.

Ubiquitin (A), CaM (B), CaM-NTD (C), CaM-CTD (D) and CaM-MLCK (E). The structures are shown in cartoon representation and colored according to residue index from red (N terminus) to blue (C-terminus). Figure is adopted from (Bakan & Bahar, 2009).

3.3 DISCUSSION

We presented here a detailed analysis of the conformational changes experimentally observed for three enzymes upon binding a broad range of ligands, and those predicted by simple physics-based models based on their native fold contact topology. In all three cases, the first principal mode of structural change, PC1, derived from experimental data exhibits a correlation of 0.78 ± 0.1 with a top-ranking mode (ANM1-ANM3) predicted to be intrinsically preferred by the (unliganded) protein. If, we further consider the correlation between subsets of PC and ANM modes (**Table 3.6**), we see that PC1 is accounted for with a cumulative overlap of 0.85 ± 0.06 by the top three ANM modes in all three cases. Similarly, the principal modes of structural variability observed in NMR ensembles exhibited remarkable correlations with top-ranking ANM modes.

We note that three PCs describe between 50% (Cdk2) and 80% (RT) of the structural variance observed in the datasets of enzymes. The structures display further heterogeneities beyond those along the first three PCs, specific to particular inhibitors, which would rather fall in the category of induced changes, succeeding the initial recognition driven by target proteins' intrinsic dynamics, as ilustrated in **Figure 3.14**.

The top ranking ANM modes are by definition collective modes of motions, and they are also known to be highly robust against sequence and structure variations. The strong correlation of experimentally observed structural changes with these ANM modes demonstrates the

collectivity and robustness of the structural changes undergone by these enzymes upon binding their ligands, even if the sizes of these concerted motions are small in many cases.

Table 3.6 Cumulative Overlap

		Cumulative overlap			
		PC1	PC2	PC3	3 PCs
HIV-RT	3 ANM modes	0.93	0.79	0.45	0.75
	6 ANM modes	0.95	0.82	0.84	0.81
	20 ANM modes	0.96	0.89	0.83	0.89
p38	3 ANM modes	0.81	0.84	0.57	0.75
	6 ANM modes	0.83	0.85	0.65	0.78
	20 ANM modes	0.89	0.87	0.76	0.84
Cdk2	3 ANM modes	0.82	0.27	0.59	0.60
	6 ANM modes	0.87	0.33	0.65	0.66
	20 ANM modes	0.93	0.46	0.85	0.77
Ubiquitin	3 ANM modes	0.42	0.75	0.39	0.54
	6 ANM modes	0.75	0.80	0.53	0.71
	20 ANM modes	0.85	0.84	0.70	0.80
CaM	3 ANM modes	0.81	0.88	0.78	0.83
	6 ANM modes	0.82	0.90	0.86	0.86
	20 ANM modes	0.84	0.93	0.89	0.89
CaM-MLCK	3 ANM modes	0.82	0.80	0.63	0.75
	6 ANM modes	0.86	0.82	0.78	0.82
	20 ANM modes	0.94	0.90	0.91	0.92
CaM-NTD	3 ANM modes	0.90	0.95	0.11	0.76
	6 ANM modes	0.91	0.96	0.57	0.83
	20 ANM modes	0.93	0.97	0.72	0.88
CaM-CTD	3 ANM modes	0.65	0.77	0.44	0.63
	6 ANM modes	0.69	0.80	0.49	0.67
	20 ANM modes	0.77	0.83	0.62	0.74

Above values are calculated using Equations 2.3.4 and 2.3.5.

What is the physical basis of ANM modes? These modes are purely based on native contact topology, or geometry. No specific interactions, other than the absence/presence of inter-residue contacts (equivalent to an excluded volume effect) are taken into consideration. The basic driving potential is entropic in origin, i.e., the directions of motions predicted by the ANM

are those entropically favored, where the uphill curvature away from the original energy minimum is minimal. The close correspondence with experimentally observed deformations suggests that the conformational changes undergone for ligand binding are dominated by entropic effects.

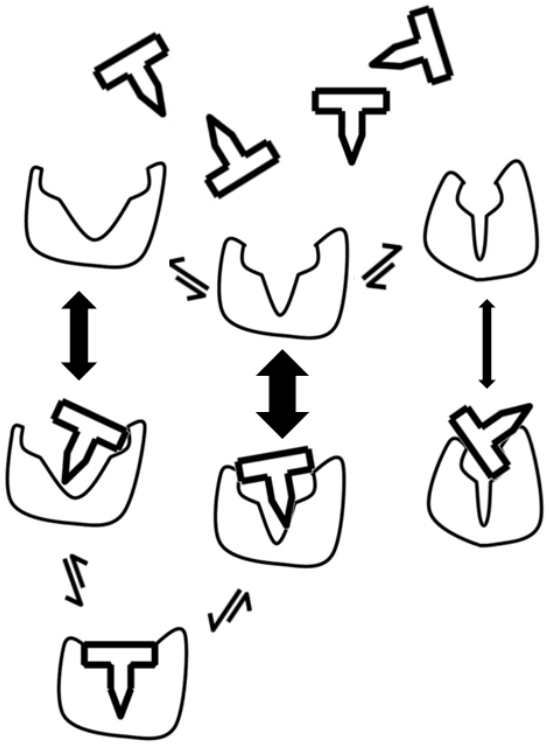


Figure 3.14 Schematic description of observed mechanism.

Inhibitor recognition and early binding events are dominated by the intrinsic dynamics of the protein. This first group of rearrangements account for 50-80% of the experimentally observed structural variance. The structural heterogeneities beyond those accounted for by top ranking PCs (or ANMs) are induced by specific ligands.

On a practical side, the fact that the structures assumed by the target proteins for recognizing their substrates comply with the global modes of motions 'predictable' by simple models such as the ANM opens the way to possible generation of representative ensembles of

conformers with the ANM. Adequate consideration of target proteins' backbone flexibility has been a major bottleneck in computer-aided drug design. A proposed solution has been to dock ligands onto multiple receptor conformations (*ensemble docking*) (Totrov & Abagyan, 2008). Sets of crystal structures, each bound to a distinct ligand (Barril & Morley, 2005), or, ensembles of NMR structures (Huang & Zou, 2007) have been considered to this aim. However, both sets may provide incomplete, if not inaccurate, information. The absence of a PCA counterpart for the ANM1s predicted for RT or for Cdk2 may signal such deficiencies. Such shortcomings are likely to be alleviated by resolving multiple X-ray structures for a given protein (Levin et al., 2007). As to NMR structures, the physical meaning of the ensemble of models deposited in the PDB - whether they represent a 'mathematical solution' that satisfies experimental (and semi-empirical) constraints, or physically accessible conformations - is yet to be established (Best et al., 2006). The correlations of the global ANM modes with the PCA modes extracted from NMR ensembles (Yang et al., 2009), consistent with present observations for ubiquitin and CaM, and with the structural variations observed in X-ray structures (Tobi & Bahar, 2005; Yang et al., 2008) are in support of the use of ANM for consolidating existing structural data, or gaining insights into potential inhibitor binding mechanisms.

4.0 TARGETING MAP KINASE PHOSPHATASES (MKPS): INSIGTS FROM STRUCTURE-BASED MODELING OF INHIBITOR INTERACTIONS

4.1 INTRODUCTION

4.1.1 Sequence, Structure, and Function

MKPs vary in length from 180 to 660 amino acids and all possess a canonical tyrosine phosphatase domain (Farooq & Zhou, 2004). Most of the members also possess an N-terminal substrate binding domain (BD) that facilitates substrate recognition and specificity.

A list of MKP catalytic domain structures resolved to date is given in **Table 4.1**. These MKPs share at least 40% sequence identity with respect to the catalytic domains of MKP-1 and MKP-3 (**Table 4.1**), in which we are particularly interested. Selected structures are shown in **Figure 4.1**, and their sequence alignments are presented in **Figure 4.2**.

Table 4.1 MKP catalytic domain structures and their sequence identities.

	Catalytic domain structures			% pairwise sequence identity among MKPs *					
Name	**PDB ID**	**Res. (Å)**	**State**	**MKP-1**	**MKP-3**	**MKP-4**	**MKP-5**	**PAC-1**	**VH3**
MKP-1	–	–	–	–	47.26	46.58	41.78	73.97	64.38
MKP-3	1MKP	2.35	Inactive	58.06	–	80.14	47.26	47.95	43.84
MKP-4	2HXP	1.83	Active	54.84	96.77	–	46.58	47.95	43.84
MKP-5	1ZZW	1.60	Active	54.84	54.84	58.06	–	42.47	34.93
PAC-1	1M3G	NMR	Inactive	77.42	51.61	48.39	48.39	–	57.53
VH3	2G6Z	2.70	Active	70.97	48.39	45.16	41.94	58.06	–

* Upper triangular entries refer to the sequence identity percentages at the catalytic domains; lower triangular entries to those at the active site region of the catalytic domain. The corresponding multiple sequence alignment is given in Figure 4.2.

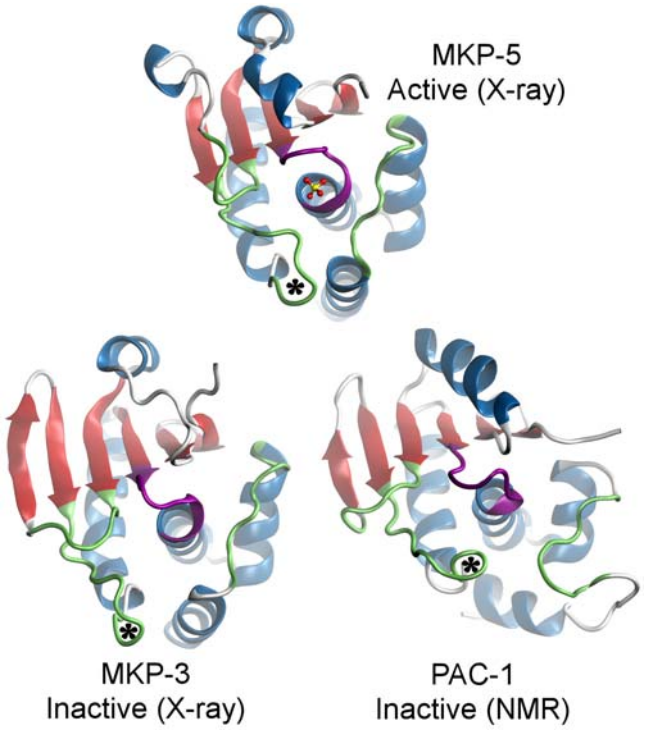


Figure 4.1 MKP CD domain structures in active and inactive states.

The active site loops display substantial changes between active and inactive states. General acid loops are marked with an asterisk. Coloring is according to the sequence alignment shown in Figure 4.2. Figure is adopted from (Bakan et al., 2008).

MKP catalytic domains share the same fold with PTPs, - a five stranded β-sheet (six stranded in the inactive state of PTPase loop) surrounded by four α-helices on one side and one α-helix on the other **Figure 4.1**. The MKP structures resolved in the inactive state of the PTPase loop, MKP-3 (Stewart et al., 1999) and PAC-1 (Farooq et al., 2003), are known to adopt the active PTPase loop conformation upon binding their substrates. On the other hand, MKP-5 (Jeong et al., 2006) and VH3 (Jeong et al., 2007) adopt the active state PTPase loop conformation even in the absence of bound substrate, consistent with the intrinsic ability of some enzymes to sample conformations that facilitate ligand binding even in the unbound state (Bahar et al., 2007). In addition, BD structures of MKP-3 (Farooq et al., 2001) and MKP-5 (Tao & Tong, 2007) have been resolved.

In the remainder of this section, the state of MKP catalytic domain in which the general acid loop is at its inactive configuration (**Figure 4.2)** will be referred to as inactive, low-activity or basal-activity state. The state in which general acid loop is at activated configuration (where it participates in the dephosphorylation reaction; **Figure 4.2**) will be referred to as activated or high-activity state.

4.1.2 Catalytic activation of MKPs upon substrate recognition offers alternative inhibition mechanisms

Substrate recognition is achieved by the BD of MKPs. In the case of MKP-1 (Hutter et al., 2000), MKP-3 (Camps et al., 1998), and PAC-1 (Zhang et al., 2005), the allosterically bound

substrate triggers the catalytic activation of the phosphatase accompanied by a movement of about 5 Å in the acidic loop (**Figure 4.1**). This geometric adjustment provides two potential inhibition mechanisms as alternatives to the conventional mechanism of competitive binding to the catalytic site. The first would involve the inhibition of substrate recognition. The second would target the obstruction of the allosteric mechanism of activation. Such an inhibitory action has been achieved for PTP1B via the restriction of a catalytically important loop from adopting an active state conformation upon binding a small molecule (Wiesmann et al., 2004).

We note that MKP-3 preferentially dephosphorylates ERK2, and in addition to the resolved structures of the catalytic domain and BD, a structure of ERK2 complexed with the kinase interaction motif from MKP-3 BD has been determined (Liu et al., 2006). The specific regions of MKP-3 that interact with ERK2 have been elucidated in systematic mutation and deletion analyses (Zhou et al., 2001); and finally the results from H/D exchange experiments (Zhou et al., 2006) have been used to construct a structural model of the MKP-3/ERK2 complex. All of these data suggest that a further examination of the binding surface and dynamics of MKP-3 is warranted and could assist in identifying novel sites for enzyme inhibition.

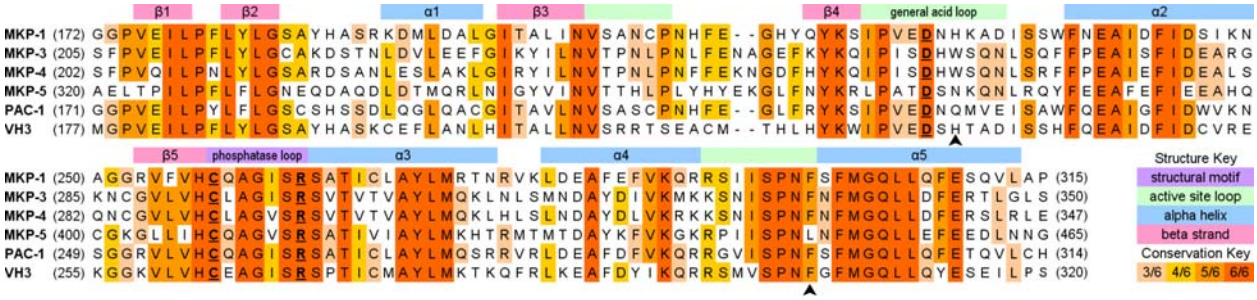


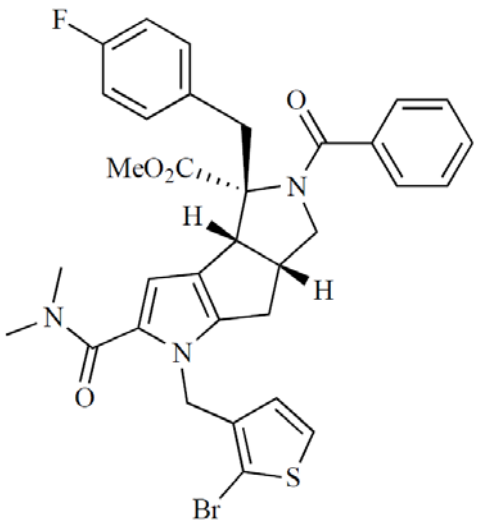
Figure 4.2 Alignment of MKP catalytic domain sequences.

Figure is adopted from (Bakan et al., 2008).

4.2 MKP-1/INHIBITOR INTERACTIONS

4.2.1 Structurally unique inhibitors of MKP-1 from a focus library of pyrrole carboxamides

The lack of readily available selective MKP-1 inhibitors has severely limited interrogation of its biological role and was one rationale for our collaborators in Dr. Lazo Laboratory to employ a recently described tricyclic pyrrole-2-carboxamide library in their screening efforts (Werner et al., 2006). In our joint report, we demonstrated the pharmacological richness of the pyrrole carboxamide library by finding that 10 of 172 members inhibited human MKP-1 (Lazo et al., 2007). One of the pyrrole carboxamides, PSI2106 (shown in **Figure 4.3**), was especially notable *in vitro* inhibitors of recombinant human MKP-1 enzyme activity with IC_{50} value (inhibitor concentration at which 50% of the enzyme activity, with respect to its maximum, is inhibited) of 8.0 ± 0.9 μM. It showed some selectivity for MKP-1 over the closely related phosphatases MKP-3, Cdc25B, VHR, and PTP1B. To understand this behavior we examined the surface properties near the catalytic site of the studied phosphatases. The compounds inhibited MKP-1 reversibly but displayed mixed kinetics. Phosphatase inhibition was retained in the presence of physiologically relevant concentrations of glutathione, - an antioxidant that quenches formation of reactive oxygen species used to show/verify that the inhibitors do not act through oxidation of catalytic cysteine. We performed detailed molecular docking studies of PSI2106 to identify key side-chains playing a role in binding these compounds.



1
PSI2106
MKP-1 IC50 = 8.0 ± 0.9 mM
MKP-3 IC50 = 43.2 ± 0.7 mM

Figure 4.3 MKP-1 inhibitor from a focused pyrrole carboxamide library.

Note the comparable activities (IC_{50} values) of this compound against MKP-1 and MKP-3. Figure is adopted from (Bakan et al., 2008).

4.2.2 Basis of specificity of inhibitors from electrostatic surface potential calculations

To assist in understanding the potential specificity of PSI2106, we examined the surface properties of the catalytic domains of the four phosphatases (**Figure 4.4**). MKP-1 and MKP-3 displayed similar surface properties in the neighborhood of the active site. In particular there was a concave region surrounded by hydrophobic residues on both structures, where the hydrophobic groups of the compounds tended to be positioned. Cdc25B and VHR, on the other hand, exhibited a significantly more polar/charged character near the catalytic site. Cdc25B active site was essentially basic, while VHR showed mostly acidic and some hydrophobic character. We

then conducted unbiased molecular docking simulations in which the 10 active compounds were docked without any pre-defined binding site to the four protein phosphatases using AutoDock (Morris et al., 1998). Thirty docking runs were performed for both enantiomers of each compound, leading to a total of 600 poses for each phosphatase. The centroids of all poses are displayed in **Figure 4.4** to identify potential binding sites where the small molecules cluster. Notably, at least three potential binding sites were predicted for the ten active compounds, one of them being near the catalytic cysteine and are seen enclosed in the blue ellipse.

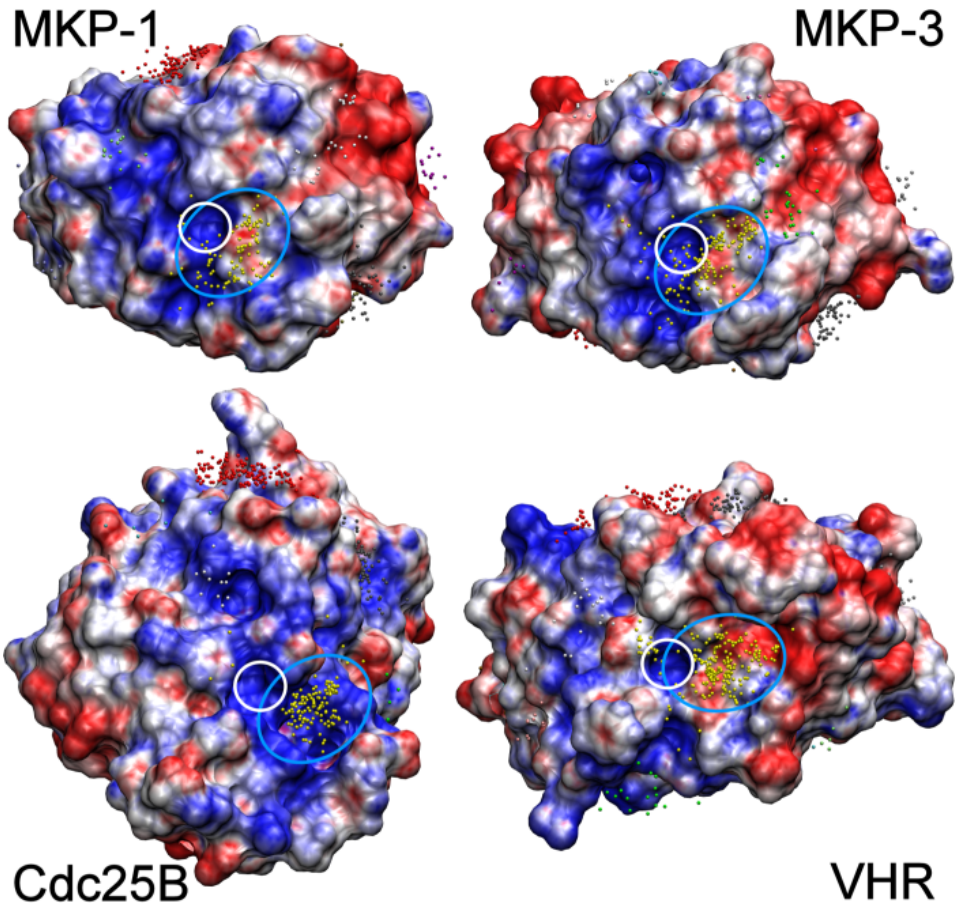


Figure 4.4 Surface properties of dual-specificity phosphatases (DSPs) and potential inhibitor binding sites. The surface representations are colored based on solvent-exposed surface electrostatic potential calculated using APBS (Baker et al., 2001). Red and blue correspond to negatively and positively charged (or polar) regions,

respectively; and white refers to neutral (or hydrophobic) regions. Each dot corresponds to the centroid of a binding pose for a compound and there are 600 hundred poses for each phosphatase. Catalytic cavities are encircled in white. The clusters in the neighborhood of the catalytic sites are colored yellow and encircled in blue. The poses in the second and third most distinctive clusters are shown in red and gray, respectively. Note the difference in the polarity/hydrophobicity of the surface within the blue circles, pointing to the origin of the differences in the selectivity of the small molecules against these DSPs. The diagrams were generated using VMD (Humphrey et al., 1996). Figure is adopted from (Lazo et al., 2007).

Docking poses were found to cluster at hydrophobic regions and large cavities on the surface. Clusters were determined using agglomerative clustering scheme and colored differently. The poses closest to catalytic sites are shown in yellow. These are proposed to form the bound conformations that can potentially exhibit competitive inhibition. Based on the comparative analysis of these binding surfaces in the four DSPs, we reasoned that the greater inhibitory action of the ten compounds for MKP-1 and MKP-3 might be associated with the hydrophobic nature of the surface near their active site.

4.2.3 Interactions of MKP-1 with pyrrole carboxamides

We further performed targeted (or biased) docking simulations in the putative binding site of MKP-1. Results from ensemble modeling of interactions were obtained using a combination of comparative modeling (MODELLER) (Sali & Blundell, 1993), structure refinement (Sybyl 7.2; Tripos, Inc. St. Louis, MO), and docking tools (GOLD) (Jones et al., 1995; Jones et al., 1997). Using the catalytic domain structures of MKPs in the active state listed in Table 4.1, we generated 300 models (conformations) of MKP-1. The objective was to take into consideration

the possible structural flexibility (and/or inaccuracy) of the modeled target proteins (encircled region in **Figure 4.5**). The two enantiomers of compound **1** were docked five times onto each MKP-1 conformation, resulting in 3,000 docking poses. Analysis of the resulting ensemble of poses to retrieve dominant patterns and identify the most favorable poses revealed the role of the solvent-exposed side-chains of His229 and Phe299 in optimizing the interactions with the inhibitor (**Figure 4.5**). In addition, a number of hydrophobic contacts involving residues Ala260, Ile262 and the β-carbons of Ser263 and Asn298 were observed. Comparable interactions were observed with MKP-3. Cdc25B and VHR, on the other hand, lacked this type of hydrophobic interactions. These residues presumably mediate the binding of pyrrole carboxamide inhibitors in favor of a geometry that occludes the access of substrates to the catalytic site.

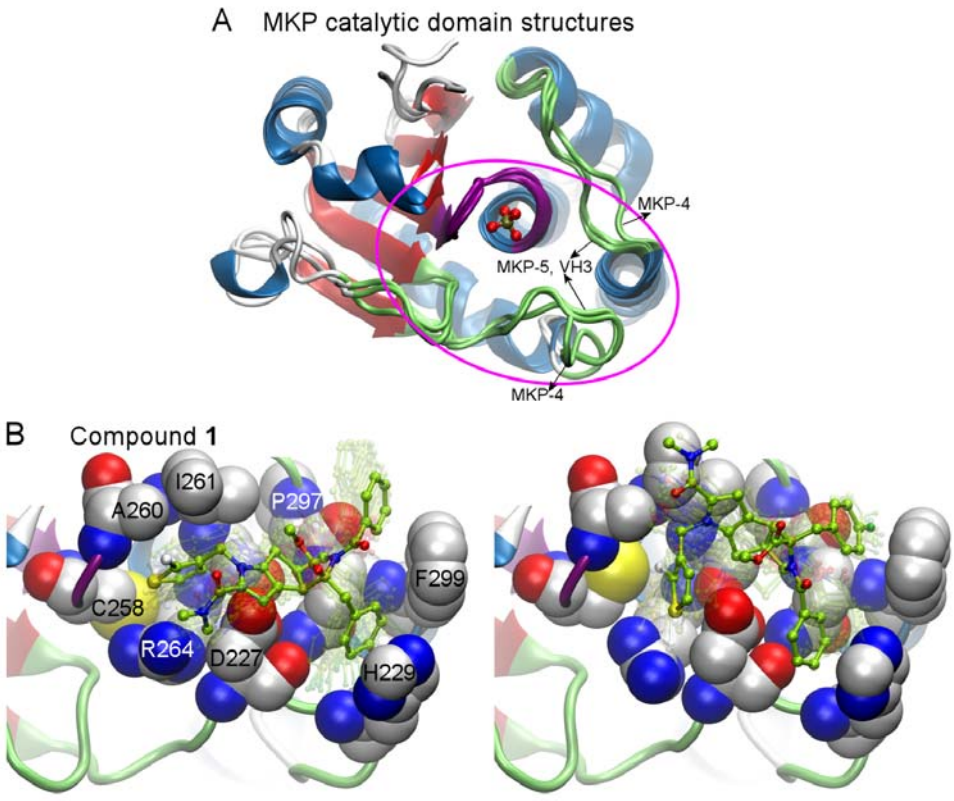


Figure 4.5 CD structures and docking solutions for MKP-1 inhibitors.

(A) Template structures used in modeling the MKP-1 catalytic domain. The inhibitor docking site is encircled. (B) Docking solutions for enantiomers of compound **1**. Note the interactions with H229 and F299, in addition to those with A260, I261 and R264 on the HCX_5R motif. Inhibitors are shown in a ball-and-stick representation with C atoms colored green. Approximately 5% of docking poses in the same bound state/cluster are shown transparently. Figure is adopted from (Bakan et al., 2008).

These observations suggest some criteria for designing MKP inhibitors. However, achieving selectivity remains a challenge. Sequence comparison between the members of the MKP family shows that MKP-1 shares a high sequence identity with others especially in the active site region (**Table 4.1**). Specifically, Phe299 is highly conserved (see **Figure 4.2**), while the His229 position appears to sustain substitutions to Trp, Asn or Glu, which may impart some selectivity. MKP-1 inhibitors are known to have comparable inhibitory potency against MKP-3 (see the IC_{50} values in **Figure 4.3**). We note that the MKP-1 His229 residue is substituted by Trp264 in MKP-3. Docking to an ensemble of MKP-3 models resulted in similar observations, in accordance with their comparable inhibitory activities. These observations highlighted the need for locating other binding sites on MKP-1 for designing selective inhibitors. We also note that these simulations were based on models generated using the active state structures of the MKP catalytic domains listed in **Table 4.1**. Further studies considering the inactive conformation of the active site loop, as well as the accessible conformations predicted by normal mode analysis, are expected to assist in generating more accurate and comprehensive predictions of MKP-inhibitor interactions.

4.3 A NOVEL MKP-3 INHIBITOR FROM ZEBRAFISH CHEMICAL SCREENS

4.3.1 Zebrafish chemical screens identify a novel MKP-3 inhibitor

Fibroblast Growth Factors (FGFs) are members of a large family of secreted glycoproteins that fulfill important functions in development, proliferation and cellular homeostasis (Thisse & Thisse, 2005). MKP-3, and other signaling proteins such as Sproutys (Spry1-4) and Sef (similar expression to FGFs) proteins, function as feedback regulators of FGF signaling. MKP-3 achieves this role by limiting the activity of ERK-1 and 2. Our collaborators identified a small molecule modulator of FGF signaling, (*E*)-2-benzylidene-3-(cyclohexylamino)-2,3-dihydro-1*H*-inden-1-one (BCI), using a transgenic zebrafish chemical screen (Molina et al., 2009) (see **Figure 4.6**). Zebrafish control experiments and *in vivo* chemical complementation assays were used to determine MKP-3 as the target of inhibitor BCI. BCI also showed comparable activity against MKP-1. IC_{50} values measured *in vivo* for MKP-3 and MKP-1 were 12.3 ± 4.0 μM and 11.5 ± 2.8 μM.

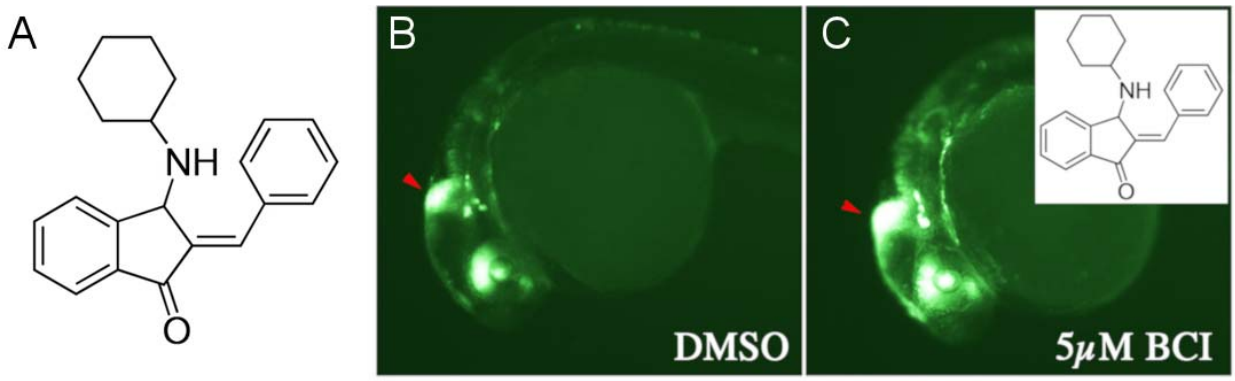


Figure 4.6 BCI structure (A) and zebrafish embryos before (B) and after (C) BCI treatment.

Panel B shows a control experiment using transgenic zebrafish embryo. GFP is tagged to MKP-3 gene (Dusp6). Panel C shows that BCI hyperactivates FGF signalling, which is measured by the increase in GFP tagged MKP-3

amount. Note that this does not necessarily mean that BCI interacts with MKP-3. Identification of MKP-3 as the target of BCI required additional experiments (Molina et al., 2009). Figure is adopted from (Molina et al., 2009).

To identify a putative binding site for BCI and to elucidate its inhibition mechanism, we performed docking simulations. We identified an allosteric binding site for BCI within the phosphatase domain. *In vitro* studies supported the model that BCI uniquely inhibits MKP-3 activation by ERK. A temporal role for MKP-3 in restricting cardiac progenitors and controlling heart organ size was uncovered with BCI treatment at varying developmental stages. This study highlights the power of *in vivo* zebrafish chemical screens to identify novel compounds targeting MKP-3, a component of the FGF signaling pathway that eludes traditional high-throughput *in vitro* screens. Here, we give the details of computational modeling and show results from *in vivo* experiments conducted to test our models.

4.3.2 Putative binding site and inhibition mechanism from modeling

A two-step process was adopted for predicting the optimal binding poses of BCI and assessing the potential mechanism of inhibition. First, unbiased docking simulations were performed where the target protein (MKP-3) was assumed to be rigid either in the low-activity state or the high-activity state. These simulations permitted us to build two hypotheses, one of which was supported by more detailed flexible docking simulations. In the following, the method and results from the two successive steps are described in more details.

4.3.2.1 Identification of potential bindings sites using unbiased docking

Crystal structures of MKP catalytic domains enabled us to perform unbiased docking simulations to identify potential BCI binding sites. BCI was docked onto two different conformations of MKP-3: the low-activity (or basal-activity) form determined by X-ray crystallography (PDB ID: 1MKP) (Stewart et al., 1999) and the high-activity form obtained by homology modeling using ORCHESTRAR (Tripos, Inc., St. Louis, MO). We used as templates the structures of MKP4 (PDB ID: 2HXP; 80% sequence identity) (Almo et al., 2007), MKP5 (PDB ID: 1ZZW; 47% sequence identity) (Jeong et al., 2006), and VH3 (PDB ID: 2G6Z; 44% sequence identity) (Jeong et al., 2007) in the high-activity state. For BCI, 400 docking poses (200 per enantiomer) were generated using AutoDock4 for each conformation (Morris et al., 1998; Huey et al., 2007). Genetic algorithm population size was set to 250. Each docking pose was selected based on the energetic evaluation of up to 5 x 10^6 alternative conformations.

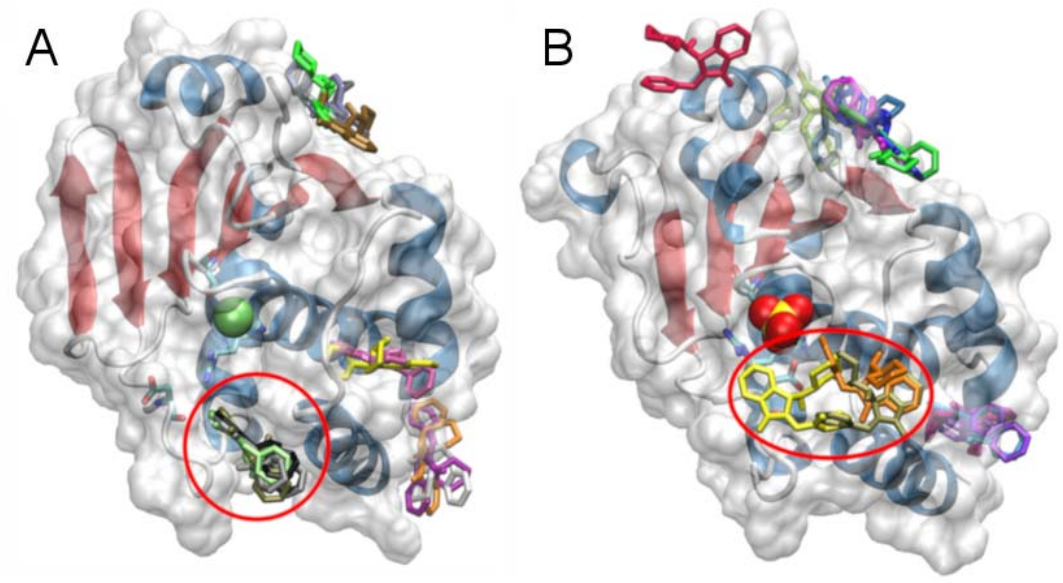


Figure 4.7 Results from unbiased docking simulations of BCI.

Results for low-activity (A) and high-activity (B) states are shown. Each BCI docking pose colored distinctly represent multiple docking solutions in a cluster. The clusters in the vicinity of the catalytic site are encircled. Clusters are determined using agglomerative clustering. The average heavy-atom RMSD (only BCI heavy atoms are used) within a cluster is less than 2 Å. Figure is adopted from (Molina et al., 2009).

The analysis of the resulting poses using an agglomerative clustering scheme revealed the clustering of a subset of binding poses in the vicinity of the active site in both conformations. In the low-activity state (**Figure 4.7A**), the binding site was a crevice known to close upon catalytic activation of the enzyme. The average AutoDock binding free energy for BCI for this site is -6.64 kcal/mol (equation 2.4.6). In the high-activity state (**Figure 4.7B**) this crevice is not accessible. Instead, a relatively more hydrophobic patch in the neighborhood of the active site was predicted to serve as an alternative binding site for the inhibitor. Predicted binding free energy for this site is -6.22 kcal/mol. Please note that the predicted binding free energies are within the standard error of most scoring functions (larger than 2 kcal/mol) (Eldridge et al., 1997; Huey et al., 2007), and they can be easily offset by small rearrangements in the protein backbone and side-chains. Due to these limitations, this unbiased docking step was used to locate potential sites for further focused docking simulations, rather than making a conclusive assessment of the binding pose and energetics. Therefore, not so much weight is usually put on AutoDock predicted energies. Also, other potential sites are shown in **Figure 4.7**. They were not considered in the first place, as BCI binding to those sites did not seem to possibly interfere with ERK binding or with the functional dynamics of the catalytic domain.

Based on predicted sites shown in **Figure 4.7** two potential inhibition mechanisms were hypothesized:

1. BCI binds the low-activity form of MKP-3 and restricts the mobility of the general acid loop so as to prevent ERK2 from inducing the conformational changes that lead to MKP-3 catalytic activation. This restricts ERK2 dephosphorylation to a basal catalytic rate.

2. BCI binds the ERK-activated MKP-3, and prevents ERK2 from optimally orienting itself, which leads to the inhibition of ERK2 dephosphorylation.

In either case, BCI was not expected to prevent MKP-3-ERK2 complex formation (at least ERK2 binding to MKP-3 binding domain) due to the large surface area of interaction distributed over two domains of MKP-3 (Zhou et al., 2006).

4.3.2.2 Flexible docking for a detailed assessment of potential inhibition mechanisms

Toward an assessment of the more likely inhibition mechanism among those hypothesized above, we further explored the binding properties of BCI by allowing the protein to undergo structural fluctuations in the neighborhood of the two above-defined states. Incorporation of backbone flexibility was important for two reasons: (i) although an X-ray structure of the inactive state is known, the general acid loop configuration is stabilized by crystal contacts in this structure as shown in **Figure 4.8**. In the absence of these contacts, this loop may assume a different conformation. (ii) The activated state of MKP-3, on the other hand, is not known. A single homology model may not represent MKP-3 activated state faithfully.

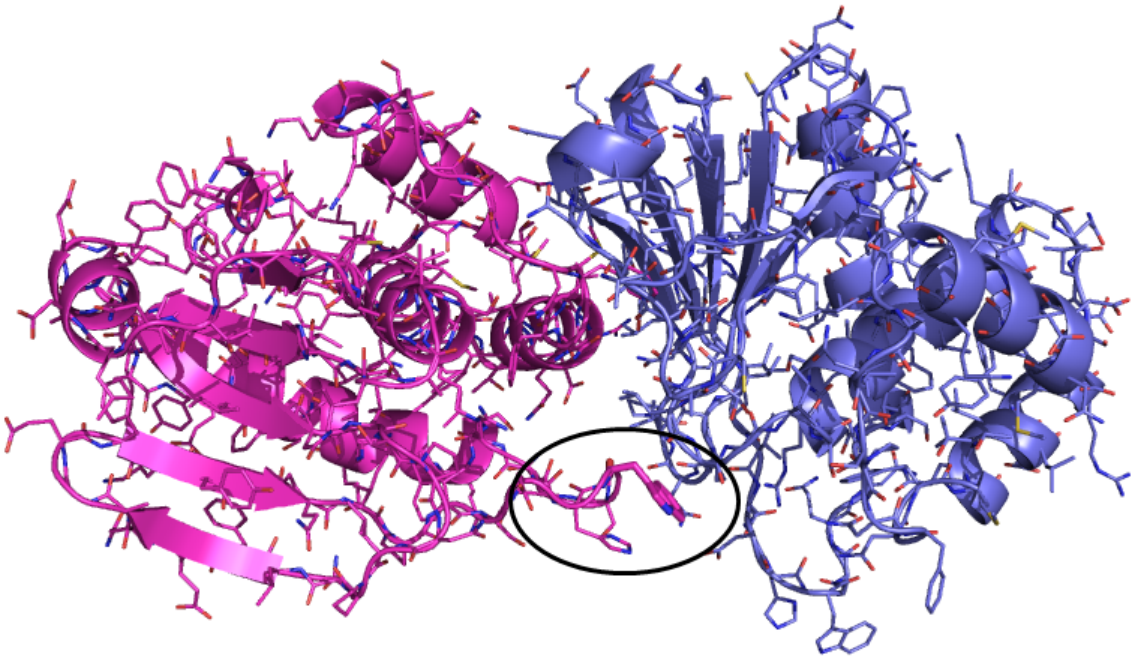


Figure 4.8 MKP-3 general acid loop (encircled region) makes crystal contacts.

Two units in the structure 1MKP resolved by (Stewart et al., 1999) are shown.

Conformations accessible near the basal-activity state were sampled by using the ANM (Atilgan et al., 2001) in combination with all-atom energy minimization. Third, fourth, and fifth ANM slow modes were found to induce a displacement in the catalytic Asp262 toward to the catalytic cavity (**Figure 4.9**) (see results at ANM web server www.ccbb.pitt.edu/anm/ (Eyal et al., 2006) using the PDB ID 1MKP) (see also Supplementary Movie 1 at http://www.nature.com/nchembio/journal/v5/n9/suppinfo/nchembio.190_S1.html). The general acid loop was also observed to have a tendency to move towards the catalytic cavity within the first 10 ns of unbiased MD simulations, in line with ANM calculations. These modes were linearly combined using scaling factors obtained from Equation 2.3.1. The high-activity state model was used as the target conformation. NAMD software (Phillips et al., 2005) and the CHARMM force field (Brooks et al., 2009) were used for energy minimization. For each α-

carbon, harmonic restraints with a force constant of 40 kcal/mole/Å^2 were defined to drive the motions along the selected ANM modes at steps of size < 0.2 Å, similar to recently introduced ANM-steered simulations (Isin et al., 2008). A total of twenty conformations (10 along each direction of the combined mode) were sampled along the selected modes by jointly optimizing backbone and side-chain conformations. As for the high-activity state, multiple models generated with MODELLER (Sali & Blundell, 1993) were used as targets.

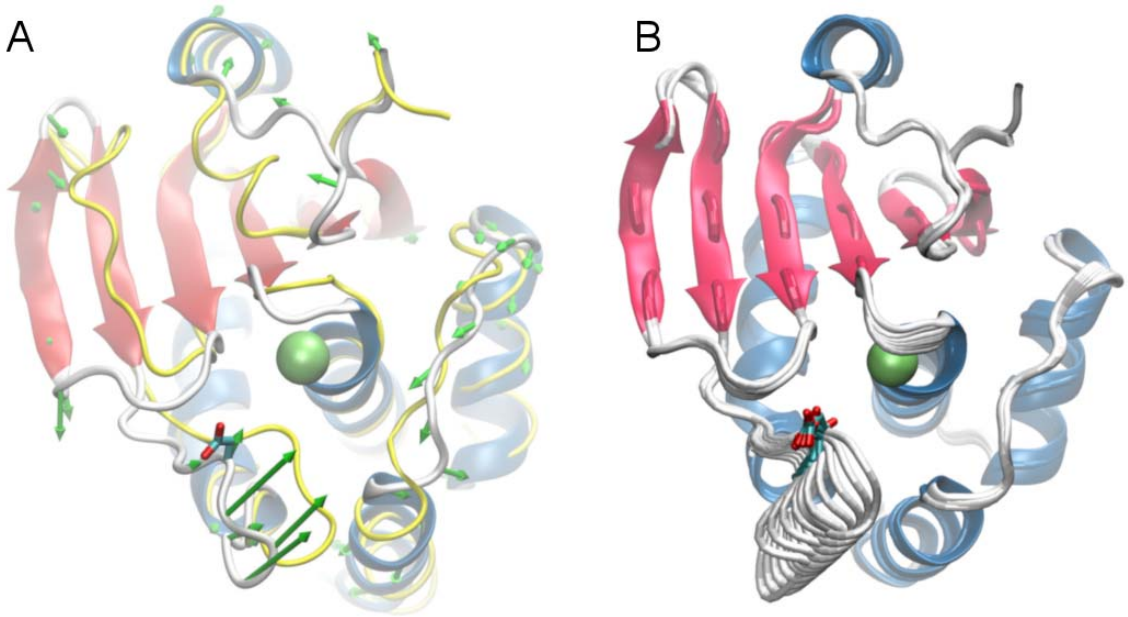


Figure 4.9 Combined ANM mode direction (A) and alternative MKP-3 conformations along this mode (B).

During docking simulations, the completely solvent-exposed side-chains have been allowed to sample rotameric states from the Penultimate Library (Lovell et al., 2000) of isomeric states. In the low activity state, Asp262, Trp264, and Asn335 were flexible. In the high-activity state, Trp264 and Phe334 were flexible, as the Asp262 and Asn335 were buried in this case. At least 1000 docking poses for the basal and activated state were generated using GOLD (Jones et

al., 1997) and cluster analysis was performed. Docking poses were scored using GoldScore (Jones et al., 1995) (see methods for details).

The most populated (and therefore entropically favorable) clusters were examined to identify the likely binding modes. The most favorable clusters (high GoldScore average and large cluster size) of BCI docking poses were located in the low-activity conformation. The most populated cluster is shown in **Figure 4.10**. The corresponding GoldScore averaged over all binding poses for this cluster was found to be 47.2 ± 2.1 (236 of 1000 poses) (see section 2.6.2 for details of this score), in favor of the hypothesized binding pose (i). In particular, the BCI molecule was predicted to preferentially fit within a crevice between the general acid loop and helix α7, rather than interacting directly with the catalytic residues Asp262, Cys293, or Arg299. At this putative binding site, a close interaction of BCI with the backbone of the general acid loop and the side-chains of Trp264, Asn335 and Phe336 was predicted (**Figure 4.10**).

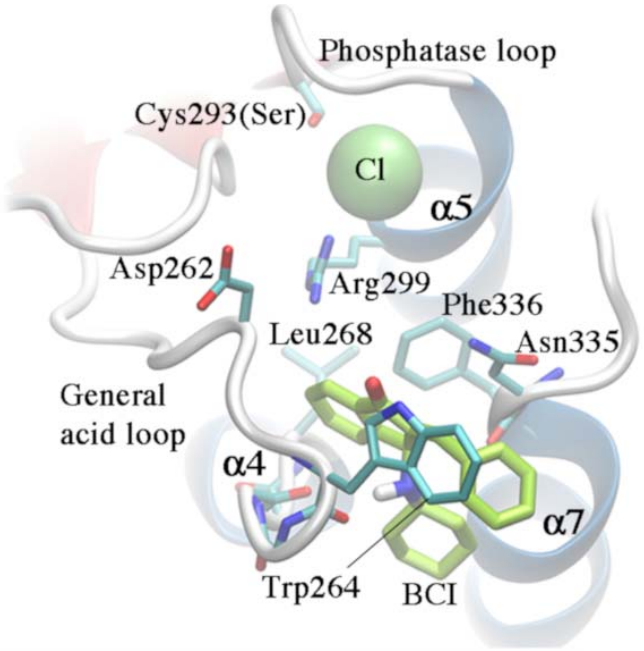


Figure 4.10 BCI interactions predicted using focused and flexible docking.

Figure is adopted from (Molina et al., 2009).

Finally, as a further verification, docking simulations were performed to compare the binding properties of BCI against VH3, MKP-1 and MKP-3. BCI was docked onto to the crystal structure of VH3 (PDB ID: 2G6Z) and a model of MKP-1 based on MKP-3 structure using the same procedure as described above for MKP-3. Docking to the VH3 crystal structure yielded much lower docking scores (27.9 ± 1.4) due to lack of the crevice observed in MKP-3 (**Figure 4.11B**), explaining lack of activity against this constitutively active homolog. Docking of BCI to the MKP-1 model resulted in comparable interactions but lower GoldScores (37.4 ± 2.9) (**Figure 4.11A**). This rationalizes BCI activity observed in zebrafish experiments and *in vivo* assays (Molina et al., 2009).

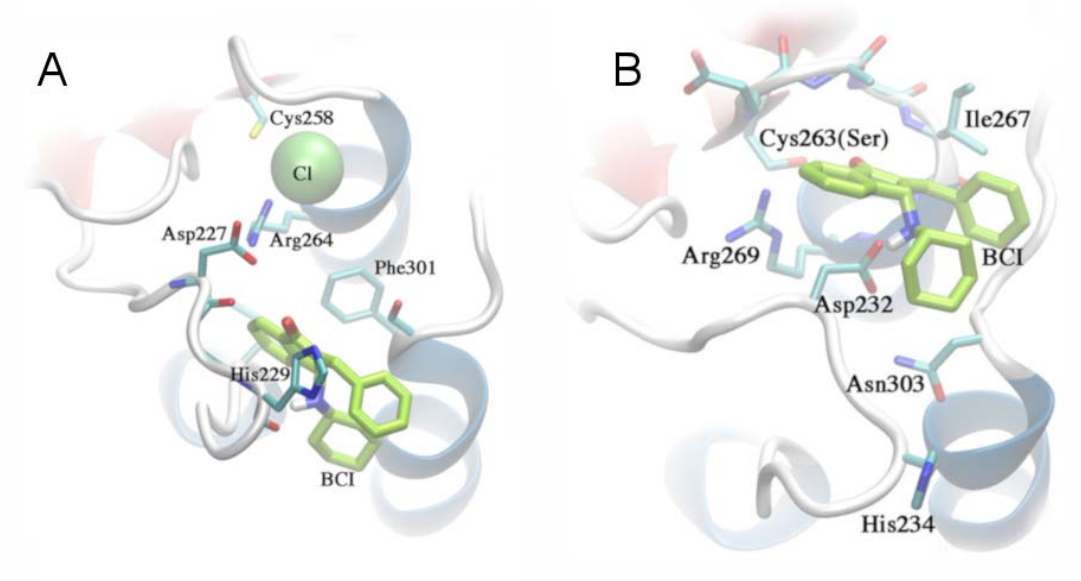


Figure 4.11 BCI interactions with (A) MKP-1 and (B) VH3.

Figure is adopted from (Molina et al., 2009).

4.3.3 Experimental testing of the hypothesis

To test these modeling predictions, we measured the dephosphorylation of a small molecule phosphatase substrate, 3-*O*-methylfluorescein phosphate (OMFP), by MKP-3 in the presence or absence of ERK2. Docking simulations predicted that BCI and OMFP could simultaneously bind within the phosphatase active site with OMFP interfacing with the core catalytic residues (**Figure 4.12A**). This suggests that BCI would not block basal MKP-3 phosphatase activity toward OMFP. Indeed, at a concentration that inhibited ERK dephosphorylation *in vitro* (100 μM, **Figure 4.12B**), BCI did not inhibit basal MKP-3 activity (Figure 4.12B). Addition of ERK2 stimulated MKP-3 dephosphorylation of OMFP three-fold and this enhancement was significantly inhibited in the presence of BCI (**Figure 4.13B**). These data support the modeling

predictions that BCI is a specific allosteric inhibitor of MKP-3 that prevents the catalytic stimulation of phosphatase activity induced by substrate binding.

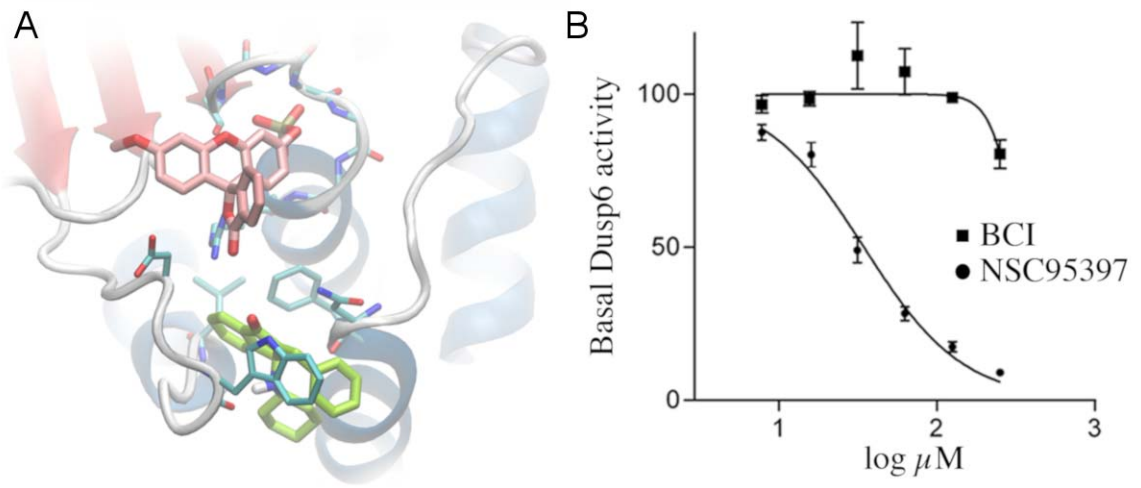


Figure 4.12 BCI does not inhibit OMFP phosphorylation.

(A) OMFP, an artificial small-molecule substrate of MKPs, and BCI docking poses do not overlap. Hence, BCI is not expected to inhibit dephosphorylation of OMFP by MKP-3. (B) Experiment showing that BCI is not effective in *in vitro* OMFP dephosphorylation experiments at concentration 20 times higher than that is effective in zebrafish experiments. BCI activity is compared to NSC95397 activity. NSC95397 is an *in vitro* active MKP inhibitor (Vogt et al., 2008). Figure is adapted from (Molina et al., 2009).

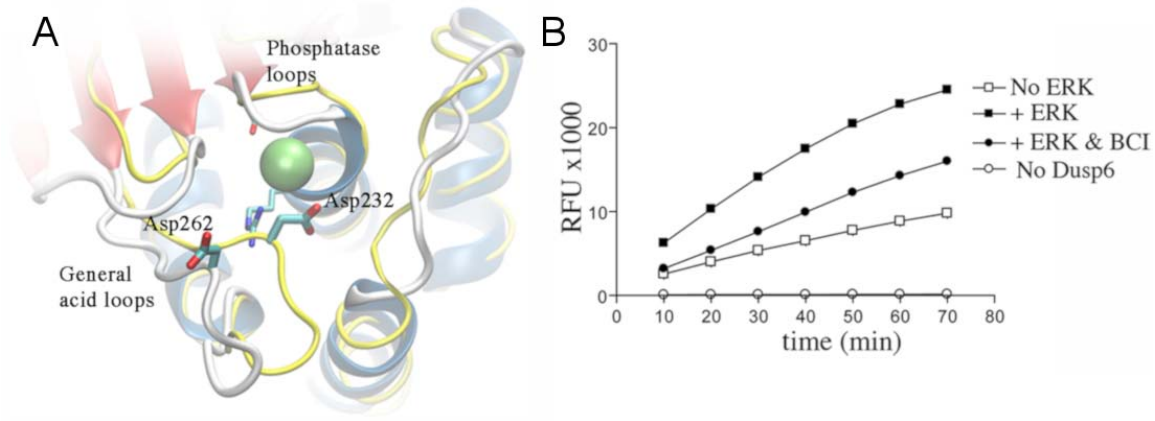


Figure 4.13 BCI is an allosteric inhibitor of MKP-3.

(A) Low-activity state MKP-3 structure is compared to high-activity state VH3 structure. When bound at the putative binding site, BCI is postulated to obstruct the flexibility of MKP-3 general acid loop so to prevent its ERK induced activation. (B) Experiments showing that BCI inhibits ERK stimulated activation of MKP-3. Addition of ERK-2 stimulated OMFP dephosphorylation by MKP-3 three-folds (raise from □ to ■) and this enhancement was significantly inhibited in the presence of BCI (decrease from ■ to ●). Figure is adapted from (Molina et al., 2009).

4.4 DISCUSSION

MKPs are therapeutically important, but challenging targets. The peculiar structural characteristics of this class of signaling enzymes have been the limiting factor of the overall progress in the field. Our molecular models have provided first insights into the interactions and selectivity of two classes of MKP inhibitors, pyrrole carboxamides and BCI. In fact in our view, the discovery of BCI and its unique MKP inhibition mechanism is the most important progress made in the field since the elucidation of catalytic activation mechanism (Camps et al., 1998) and determination of the structure (Stewart et al., 1999) of MKP-3.

In our work, we benefited from 'ensemble analysis' approach. Standard tools and approaches, such as using a single target conformation or evaluating docking poses solely based on scoring function, would have held back our progress. Considering alternate MKP conformations in docking and ensemble (clustering) analysis of docking poses enabled us locating putative binding sites and elucidating inhibition mechanisms of MKP inhibitors.

A limitation of our models is partly due to the small range of experimental IC_{50} values that MKP inhibitors exhibit. In section 4.2, we showed results for only one pyrrole carboxamide, PSI2107. In our published work, we discussed experimental results for 9 more notable pyrrole carboxamides from the larger screening library of 172 compounds (Lazo et al., 2007). Docking poses for these compounds can be found at http://www.ccbb.pitt.edu/cadd/open/MKP/JPET/. The IC_{50} values of these compounds range from 8 to 19 μM. In addition to these, the compounds with lowest measured activity had IC_{50} values of 30 μM. Hence, the difference in the affinities of the best (PSI2107) and worst compounds in this library is much smaller than the standard error of most docking scoring functions (scoring functions cannot rank order compounds with maximum binding affinity difference less than 100 folds). This has limited us to understanding the selectivity of these compounds and identifying residues potentially important for binding. A 3-dimensional structure activity relationship that can be predictive of activities of new compounds could not be developed.

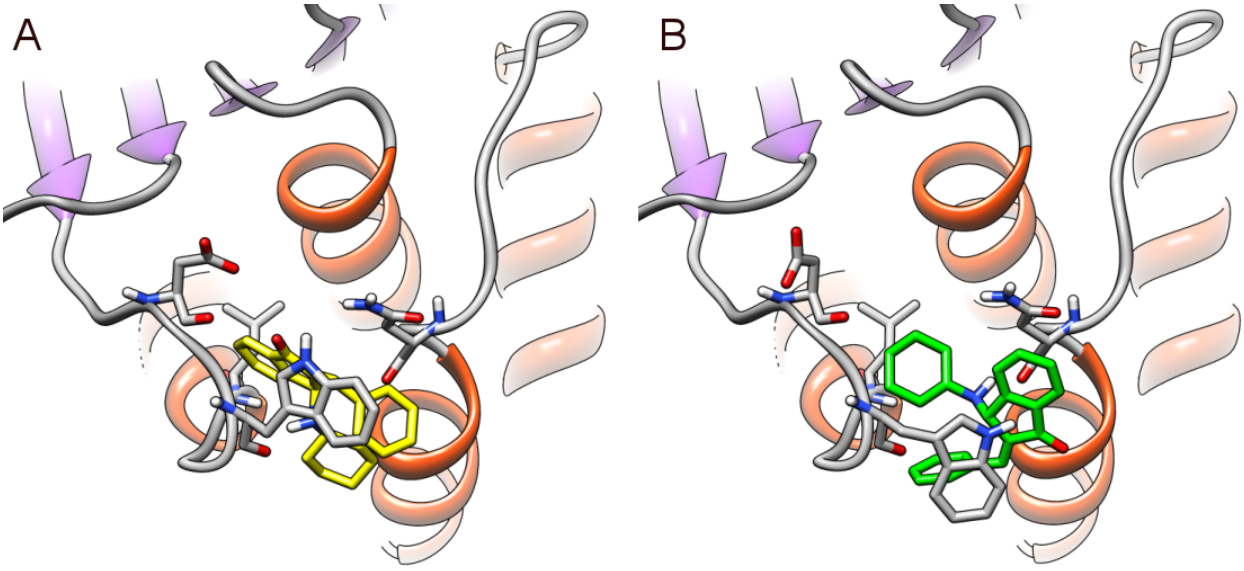


Figure 4.14 Multiplicity of BCI binding modes.

Alternate BCI binding modes with comparable docking scores are shown. Gold score average for (A) is 47.2 ± 2.1 (236 poses) and for (B) is 46.1 ± 2.9 (156 poses).

Another limitation arises from the weaknesses of typical scoring functions. Although our BCI interaction model is supported by novel *in vitro* experiments, we have multiple likely binding modes at the putative binding site. Two alternate configurations of BCI, for one of its enantiomers, are shown in **Figure 4.14**. This multiplicity is further complicated by the chirality of BCI, which exists as two enantiomers. Not being able to single out the most populated BCI configuration, prevents us from suggesting modifications on BCI structure to improve its potency. Hence, the design of compounds based on BCI is yet a trial and error procedure. Once the more potent enantiomer of BCI and its most populated configuration is identified, it will enable the rapid development of more potent and possibly selective analogs. One of the ways to achieve structure-based design of BCI analogs is of course to obtain structural, preferably X-ray crystallographic, data on MKP-3 and BCI complex. An alternate way is an iterative design-synthesis strategy, whereby new computational hypotheses can be evaluated by specifically prepared synthetic small molecules. This will be discussed in the last section of this work.

5.0 UNDERSTANDING FUNCTIONAL MECHANISMS OF MEMBRANE PROTEINS

5.1 INTRODUCTION

As we have discussed above, understanding the interactions between a target protein and its inhibitors is of crucial importance in drug discovery (Congreve et al., 2005). Molecular docking is the primary computational tool to model these interactions (Brooijmans & Kuntz, 2003) and screen compound libraries of small-molecules with potential inhibitory/agonistic/antagonistic activities (Shoichet et al., 2002). There are numerous successful applications of docking to membrane proteins. Predix Pharmaceuticals, for example, targeted five different GPCRs in *in silico* screens of commercially available libraries and identified 11 compounds per target, with an average hit rate of 17% (Becker et al., 2004). In another study, Wang and coworkers targeted dopamine (D_3) receptors, and identified four compounds that bind at 100 nM levels, with 60% hit rates (Varady et al., 2003).

The ligand-selective conformational heterogeneity of GPCRs has been recognized, however, as a limiting factor in *in silico* efforts (Kenakin, 2003; Kobilka, 2007). The binding site geometry of GPCRs differ, depending on the functionality and the potency of bound ligands (Ghanouni et al., 2001). Kinetic measurements and single molecule spectroscopy both reveal that

the 7 trans-membrane (TM) helix bundle samples distinguishable conformational states in the absence or presence of ligand, and the populations of these conformational states shift upon ligand binding (Peleg et al., 2001; Swaminath et al., 2004). State-of-art docking programs usually allow for only partial binding site flexibility limited to optimizing a small number of side-chain rotations or short loop conformers. Overlooking backbone conformational flexibility hampers the success of *in silico* drug discovery.

Abagyan and coworkers made prominent contributions to developing algorithms and tools that take account of target protein conformational flexibility (Totrov & Abagyan, 2008; Bottegoni et al., 2009), which have been successfully applied to GPCRs (Cavasotto et al., 2003). In particular, a ligand-steered homology modeling approach was developed, which uses existing ligands to shape and optimize the GPCR's binding site (Cavasotto et al., 2008). The idea therein is to start with hundreds of crude homology models as probable conformations of the target protein, then filter them based on their interaction energy with known ligands probed by flexible docking and on their ability to detect known ligands in virtual screening tests. The utility of this approach was demonstrated by its application to melanin-concentrating hormone receptor 1 where a 10-fold improvement over random high-throughput-screening was achieved and six novel antagonists were identified. In a similar recently published study, β_2AR interactions with agonist/antagonist were examined upon generating multiple conformations of β_2AR (Reynolds et al., 2009). The models were reduced and further refined by flexible docking of selected agonists in the light of mutagenesis data to obtain models that outperformed rhodopsin-based models. In accord with these findings, Kobilka and coworkers reported that rhodopsin-based homology models of β_2AR developed prior to β_2AR structure resolution were more similar to rhodopsin

rather to β_2AR (Cherezov et al., 2007), stipulating the need to consider more distinctive target conformations.

The generation of multiple conformations for the target protein emerges as an important component of computational tasks for modeling and simulating membrane protein-inhibitor interactions. NMA with ENMs appears to be particularly suitable for generating backbone rearrangements. It suffices to have but one structure to generate a distribution of energetically favorable conformations in its neighborhood. Likewise, the method can be used to refine/broaden an existing collection of conformations. Here, we illustrate the utility of ANM for sampling of alternate conformations of rhodopsin and nAChR and review the relevant literature. Our results show that ANM is able to capture robust collective modes of motions of not only soluble proteins, but also membrane proteins, despite the changes in the environment due to the lipid bilayer. Hence, we can say that functional motions correlate with intrinsic dynamics of the proteins which can be represented using simple physics-based models.

5.2 RHODOPSIN

5.2.1 G-protein couple receptors

G-protein coupled receptors constitute one of the largest protein superfamilies in the human genome with more than 800 members. Among the five families that form this superfamily, the rhodopsin family is the largest, with 701 members (Fredriksson et al., 2003). All GPCRs share a

common architecture of seven TM α-helices (H1-H7) (**Figure 5.1A**). They transmit EC signals to the CP region via allosteric movements of TM helices. The resulting changes in the CP surface facilitate G-protein binding and activation, which, in turn, triggers a cascade of cellular responses (Kobilka, 2007; Oldham & Hamm, 2008).

5.2.2 Rhodopsin

The vast majority of the structure-based computations for GPCRs have been done using the bovine rhodopsin structure, originally resolved by Palczewski and coworkers (Palczewski et al., 2000). In addition to the bundle of seven TM helices, referred to as opsin, the structure contains an 11-*cis*-retinal (chromophore) deeply embedded in the core (**Figure 5.1A**). The EC domain consists of the N-terminus and three inter-helical loops EC1-EC3; the CP domain contains three inter-helical loops CL1-CL3 connecting respective pairs of helices H1-H2, H3-H4 and H5-H6, and a C-terminal helix H8 that runs parallel to the membrane. The EC domain contains a β-sheet, which serves as a lid to the chromophore binding pocket, stabilized by a highly conserved disulfide bond between Cys110 and Cys187. The retinal, covalently bound Lys296 on H7, undergoes a *cis/trans* isomerization upon light activation. This gives rise to a local conformational strain that propagates through the concerted rearrangement of the TM helical bundle to the CP domain, inducing an opening at the conserved D(E)RY motif (residues shown in sphere representation in **Figure 5.1**), which is recognized by the G-protein (**Figure 5.1B**). The active form, metarhodopsin II, is reached after a series of photointermediates. It binds the heterotrimeric G-protein, transducin, and interacts with several other signaling proteins.

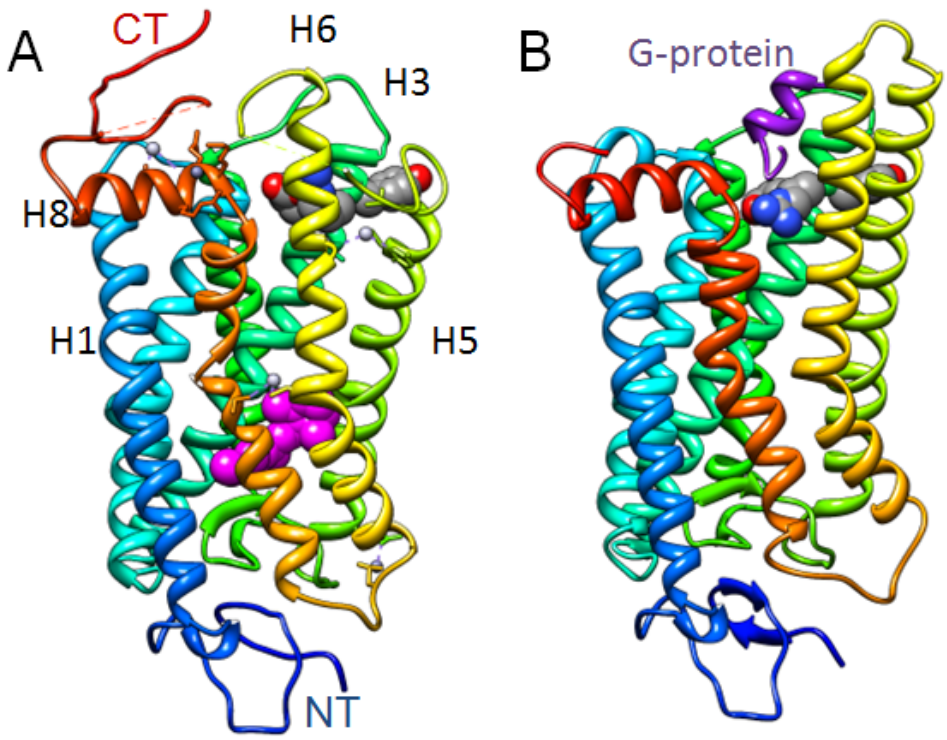


Figure 5.1 Rhodopsin (A) (chromophore in magenta) and G-protein bound opsin* (B).

Recent years have witnessed a remarkable progress in the number of newly solved GPCR structures (Hanson & Stevens, 2009). Comparison of the structures of bovine opsin in its G-protein-interacting form (referred to as opsin*) (Scheerer et al., 2008) and rhodopsin shows, for example, an outward tilt of 6 Å in TM6, and pairing of TM5 to TM6. Comparison of the ligand-free opsin (Park et al., 2008) and opsin*, on the other hand, shows little structural difference, suggesting that the opsin conformational population is shifted towards the activated state in the absence of retinal and G-protein.

The type and extent of conformational changes undergone upon activation of rhodopsin have been extensively examined by various experiments (Resek et al., 1993; Farrens et al., 1996; Altenbach et al., 1996; Hubbell et al., 2000; Altenbach et al., 2001a; Altenbach et al., 2001b;

Hubbell et al., 2003; Kusnetzow et al., 2006; Bhattacharya et al., 2008; Altenbach et al., 2008) and computations (Rohrig et al., 2002; Saam et al., 2002; Crozier et al., 2003; Rader et al., 2004; Shacham et al., 2004; Huber et al., 2004; Lemaitre et al., 2005; Isin et al., 2006; Grossfield et al., 2007; Isin et al., 2008; Bhattacharya et al., 2008). GNM and ANM studies (Rader et al., 2004; Isin et al., 2006) show that the global mode is controlled by a broad hinge-bending region that includes the chromophore binding pocket and a number of highly constrained conserved residues in the close neighborhood such that the structural changes locally induced upon the isomeric transition of the *cis*-retinal are efficiently propagated through cooperative rigid-body movements of the TM helices, towards both the CP and EC regions. An effect of these cooperative movements is opening the CP ends of the TM helices 3, 4 and 6, thus exposing the ERY motif at the G-protein binding site. A model for the Meta II state has been proposed by analyzing the lowest ANM modes in conjunction with experimental data (Isin et al., 2006). The model was shown to correctly predict 93% of the experimentally observed effects in 119 rhodopsin mutants for which the decay rates and misfolding data have been reported, including a systematic analysis of Cys to Ser mutations.

5.2.3 PCA of rhodopsin structure ensemble

With the elucidation of a large number of structures, we are now in a position to examine more closely the correlation between the experimentally observed structural differences and theoretically predicted conformational changes. We performed a PCA of currently available rhodopsin and opsin structures, and compared the resulting PC modes to ANM modes. Our

dataset includes 16 structures, comprised of 14 rhodopsin and two opsin X-ray structures. Out of N = 348 residues, 312 are commonly resolved in the dataset of examined structures, excluding the segments 230-240 on CL3, 311-313 between H7 and H8, and 327- 348 at the C-terminus. The distribution of the structures along the first two principal modes is shown in **Figure 5.2**. These two modes contribute about 62% and 12% respectively, to the structural variability in the dataset. The PCA clearly separates the structures into two clusters along the first principal axis. These two clusters may be viewed, in a sense, as the two substates illustrated in **Figure 1.2**. Notably, the first cluster includes all the 14 rhodopsin structures in the inactive (sub)state, and the second, two opsin conformations in the putative active (sub)state. Mode 1 therefore unambiguously distinguishes between these two substates, representative conformations of which are displayed in **Figure 5.3A**. The 2nd principal mode, on the other hand, further disperses the structures within the first cluster. This mode essentially refers to the changes in loop conformations and termini orientations. These can be viewed as the microstates in the inactive substate.

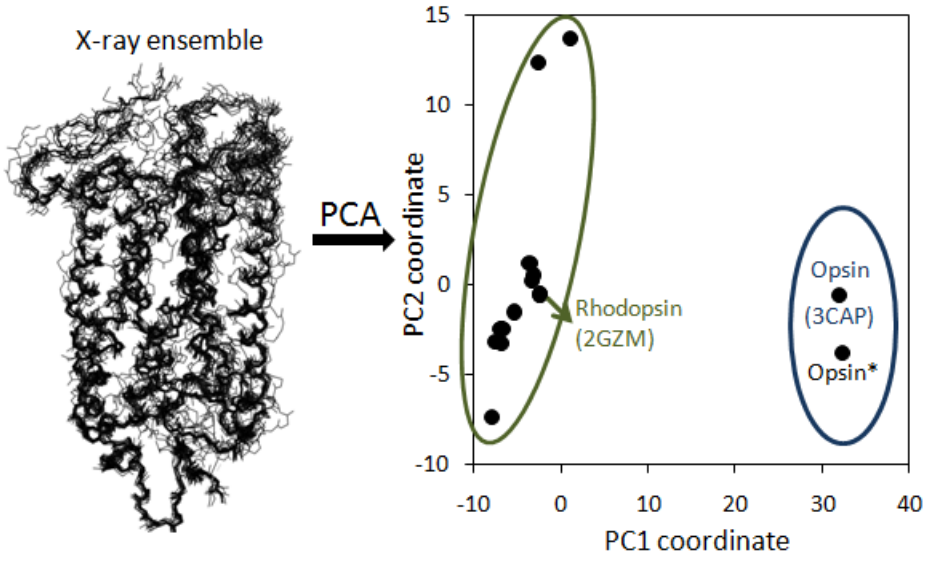


Figure 5.2 PCA of rhodopsin structural ensemble.

Distribution of 16 X-ray structures in the subspace spanned by the PCA mode directions 1 and 2. These respective modes account for 62 % and 12% of the structural variability in the dataset. The principal axis 1 differentiates the inactive and (putative) activated structures which are clustered in two distinctive groups, and the PCA axis 2 further differentiates between the structures in the cluster of inactive rhodopsins.

5.2.4 Correspondence between ANM modes and PCA modes

Panel B in **Figure 5.3** illustrates how the rhodopsin conformation (green) is closely reproduced upon reconfiguring the opsin structure along $\boldsymbol{p_1}$. Comparison of the range of the principal axes 1 and 2 in **Figure 5.2** shows that the size of motions along $\boldsymbol{p_1}$ is at least twice as large as that along $\boldsymbol{p_2}$. Thirdly, and most importantly, the principal modes may be directly compared with those predicted by NMA. The PCA modes are exclusively based on experimental data for an ensemble of structures, while ANM modes are predicted by the theory for a single structure. Comparison of the two sets can help benchmark the computational predictions provided that the experimental

dataset represents a more or less complete ensemble (see for example the study performed by Jernigan laboratory for HIV-1 protease), or consolidate the results, given that both sets involve approximations. In the present case, the set of PDB structures is far from complete. Yet, ANM calculations performed for the two representative structures (labeled) from each cluster showed that $\boldsymbol{p_1}$ exhibits a cumulative overlap of 0.79 with the first 20 ANM modes intrinsically accessible to opsin; and a cumulative overlap of 0.74 with the first 20 ANM modes accessible to rhodopsin. Thus 2% of 930 ANM modes from the low frequency regime provide a reasonable description of the change observed experimentally. The reconfiguration predicted by moving the opsin structure along these ANM modes is shown in **Figure 5.3C**. These results again confirm the view that the relative movements of the TM helices 5 and 6 observed upon light activation are intrinsic properties encoded in the rhodopsin architecture.

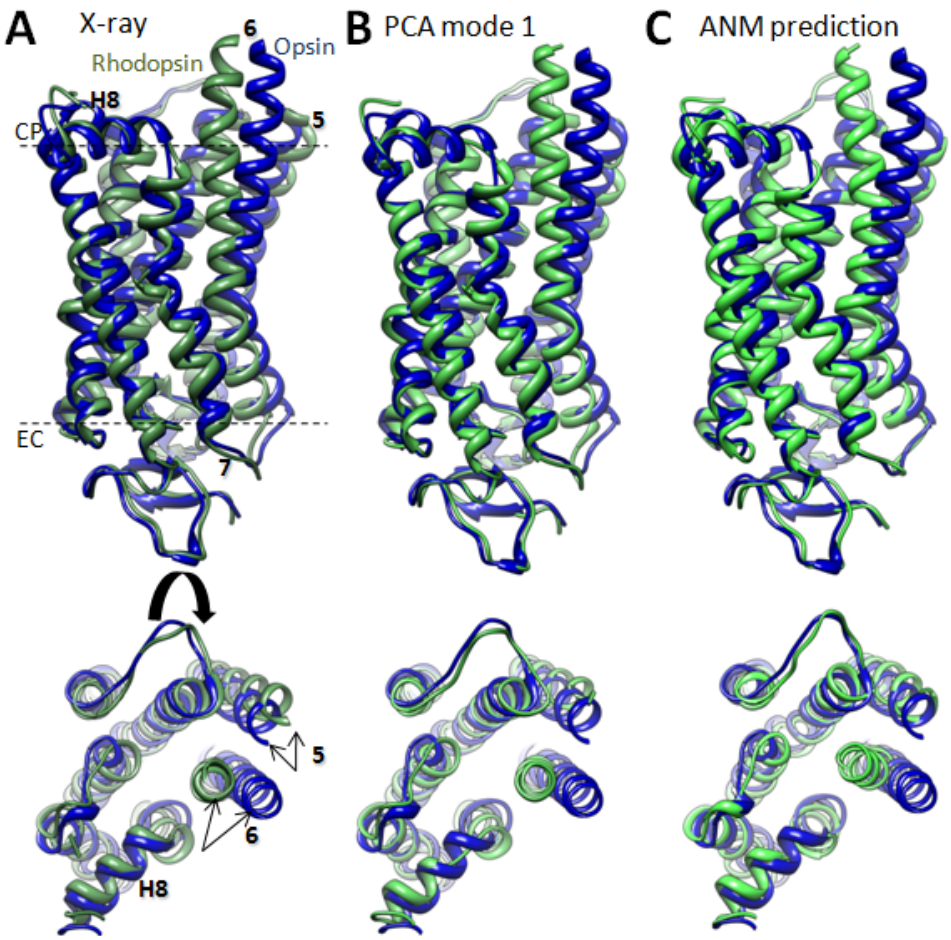


Figure 5.3 Opsin structures deformed along PC and ANM modes.

(A) Superimposition of experimentally determined rhodopsin and opsin structures. (B) Rhodopsin structure generated by deforming the opsin structure along the first principal mode, p_1. (C) Rhodopsin conformation predicted by deforming the opsin structure along the 20 lowest frequency ANM modes. Structural models in B and C contained C_α atoms, only; the remaining backbone atoms were reconstructed with BioPolymer module of Sybyl 8.3 (Tripos). ANM calculations were performed using the relatively short cutoff distance of R_c = 8 Å, so as to release interhelical constraints.

5.3 NICOTINIC ACETYLCHOLINE RECEPTOR

5.3.1 Ligand-gated ion channels

Communication between nerve cells takes place at junctions called synapses. The presynaptic cells release, upon activation, neurotransmitters into the synapse, which bind to ligand-gated ion channels (LGICs) on the surface of the post-synaptic cells. Binding of neurotransmitter causes the channels to open, allowing the ions to flow across the postsynaptic-cell membrane. The opening and closing of LGICs rapidly convert chemical signals into an electrical output, regulating the flow of information. Mutations in LGICs lead to a number of 'channelopathies' such as congenital myasthenic syndromes, epileptic disorders and hereditary hyperekplexia (Czajkowski, 2005). Approximately 8.3% of small-molecule drugs target LGICs (**Figure 1.5**).

5.3.2 Nicotinic acetylcholine receptor (nAChR) structure

Nicotinic acetylcholine receptor is a member of a superfamily of pentameric transmitter-gated ion channels, also called Cys-loop receptors, which include the serotonin 5-HT3, GABA A and GABA C, and glycine receptors. Members of this superfamily contain a signature loop of 13 residues closed by a disulfide bridge, called Cys-loop, at the interface between the EC and TM domains of their respective monomers (Sine & Engel, 2006). The nAChR activity is triggered by binding of acetylcholine (ACh), or nicotine.

The structure of nAChR in the closed state has been determined by cryo-EM of tubular crystals grown from the electric organ of *Torpedo marmorata* (Miyazawa et al., 2003; Unwin, 2005). The structure consists of five subunits (α, β, δ, α, and γ), two of which (α-subunits) have a slightly distorted (closed or tense, T) conformation compared to the other three (open, relaxed, R), hence the pseudo-symmetric organization of the quaternary structure. The receptor is organized into three domains (**Figure 5.4A**): a large N-terminal EC domain involved in binding the neurotransmitter, a TM pore domain, and a smaller CP domain. The N-terminal domain of each subunit is composed of an N-terminal α-helix and two β-sheets arranged in a curled β-sandwich connected by the Cys-loop (**Figure 5.4A**). The same fold is exhibited by the soluble ACh binding protein (AChBP) (Brejc et al., 2001). There are two ACh binding sites at the interfaces between the α-δ and α-γ subunits' EC domains. The TM domains of individual subunits are composed of four helices, M1-M4, overall forming a cluster of 20 TM helices. The pore lining helix, M2, is tilted radially inwards towards the central axis up to the middle of the membrane. The outer helices (M1, M3 and M4) tilt both radially toward, and tangentially around, the central five-fold axis (Miyazawa et al., 2003). Comparison of the ligand-free nAChR and ligand-bound AChBP structures suggests that ACh binding induces a local structural rearrangement (closure of two loops around ACh) to convert the α-subunits to their open (relaxed, R) state, which cooperatively triggers a transient opening of the channel pore at a distance of about 40 Å, thus allowing cations, particularly Na^+ and K^+, to pass through.

5.3.3 Models of nAChR channel gating

Several models have been proposed for elucidating the gating mechanism of nAChR (Changeux & Edelstein, 1998; Hung et al., 2005; Taly et al., 2005; Cheng et al., 2006; Szarecka et al., 2007; Liu et al., 2008). NMAs performed by different groups for the complete structure of nAChR and for the EC-TM domains of the homopentameric α_7 nAChR models based on the nAChR and AChBP structures (Taly et al., 2005; Cheng et al., 2006) invariably showed that the lowest frequency mode is a concerted *quaternary twist* with counter-rotations of the EC and CP domains around the five-fold symmetry axis. We performed ANM calculations for the nAChR structure with PDB ID 2BG9 (Unwin, 2005). The softest ANM modes directions are displayed in **Figure 5.4A**. Like all vibrational modes, this global mode gives rise to two sets of conformers, corresponding to positive and negative movements along the mode axis, manifested as opposite torsions in this case. Of these two sets, one is found to induce an opening in the TM channel of nAChR: the counterclockwise torsional rotation of the TM domain accompanied by clockwise rotation of the EC domain when viewed from the CP region. As can be seen in **Figure 5.4B**, the five M2 helices lining the pore are displaced slightly away from the center during this particular quaternary twisting. The calculation of the pore size profile along the TM channel (using HOLE (Smart et al., 1996)) shows that a relatively small (up to ~ 3Å) increase in diameter is induced in the constriction zone, the original value of which is 5.7 Å in the known structure. The diameter of the first hydration shell of a monovalent cation is typically around 8 Å. This small opening of the pore induced by the global mode is thus expected to enable the passage of hydrated cations (Taly et al., 2005).

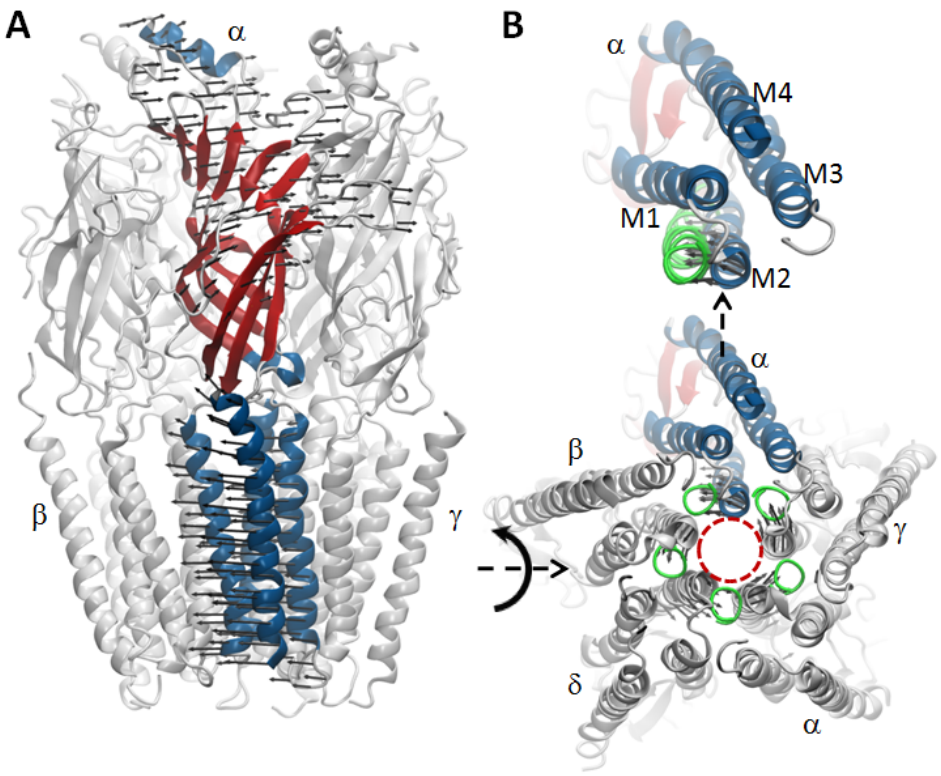


Figure 5.4 Ligand-gated ion channel nAChR structure and dynamics.

(A) Structure of the EC and TM domains of nAChR (PDB ID: 2BG9). The secondary structure of one of the monomers (α) is colored to display the β-sandwich fold (red) of the EC domain, and the four TM helices (M1-M4; blue) of the TM domain; and the remaining four monomers are shown in gray. The lowest frequency ANM mode induces a quaternary symmetric twist as indicated by the arrows shown for monomer α. (B) CP end of TM domain (bottom) and close up view of one of the monomers (monomer α, colored) (top). Red dashed circle indicates the channel pore. Arrows indicate the collective movements of M2 helices along ANM mode 1. Green circles represent the CP end of the M2 helices after deformation along ANM mode 1. Figure is adopted from (Bahar et al., 2009).

An increase in the pore radius by ~1.5 Å has indeed been suggested by MD and Brownian dynamics simulations to be sufficient to raise the computed conductance to ~ 22 pS, - a value comparable to the experimental measurements for the open channel (Corry, 2006). The above results from NMAs (including those obtained with ENMs) support the view that small but

concerted rearrangements of the M2 helices lining the pore readily allow for an expansion of this size in the pore, thus providing an efficient gating mechanism (Unwin, 1995). Concerted rigid-body motions of M2 helices were inferred by Unwin from early comparisons of the original structures at various resolutions. Grosman and coworkers made extensive single-channel electrophysiological measurements to analyze the change in the microenvironment of the helices M1, M2 and M3 between the open and closed forms of the channel (Cymes et al., 2005; Cymes & Grosman, 2008). Mainly, they examined the position-dependent proton transfers (or pKa shifts) for ionizable residues that have been engineered in the inner faces of these helices. These experiments led them to conclude that nAChR pore dilation involved subtle rearrangements, if any, of these three helices (Cymes et al., 2005; Cymes & Grosman, 2008). Notably, the twisting mode predicted by the NMA does not necessarily implicate any significant change in the orientation of the M2 helix side chains with respect to the channel lumen, but small rotations of about ~15° that presumably induce minimal changes in the exposure of side chains, which may explain the experimental observations. The changes induced by the NMA-predicted quaternary twisting mode, in the exposure of M2 residues' side chains to the central pore, were indeed pointed out by Changeux and coworkers to be compatible with the data from Grosman and coworkers' experiments (Taly et al., 2005).

The global *twisting-to-open* motion of nAChR resembles those observed in other multimeric ion channels (Bahar et al., 2009). The collective modes of the M2 bundle (pore-lining helices) predicted by NMA are also observed in PCA of MD simulation trajectories (Hung et al., 2005). Conventional MD simulations of 30ns for nAChR embedded in explicit lipid bilayer also indicate (Liu et al., 2008) the concerted rotations of M1 and M2 helices accompanying the

shrinking of the ACh binding pocket, and the open-close transition of the structure can be driven by introducing a torsional rotation around the pore axis in steered MD. The accord between NMA results for the nAChR in the absence of lipid environment, and MD trajectories conducted in explicit water and lipid bilayer, corroborate the dominant role of the membrane proteins' intrinsic features in defining the movements that facilitate essential functions such as gating.

5.3.4 Recent structures confirm quaternary twist-to-open mode

The recently resolved X-ray structures of two bacterial homopentameric ligand-gated ion channels shed further light into pore opening/closing mechanisms. These are the closed state structure of *Erwinia chrysanthemi* ligand-gated ion channel (ELIC) (Hilf & Dutzler, 2008) and two open-state structures of *Gloebacter violaceus* ligand-gated ion channel (GLIC) (Bocquet et al., 2009; Hilf & Dutzler, 2009). These structures do not include the CP helical bundle, but bear EC and TM domains comparable in size and fold to their counterparts in nAChR. In particular their EC domain superimposes closely with AChBP and with the EC domain of nAChR, except for a missing α-helix. The most striking difference between the ELIC and nAChR structures is at their pore domain: the EC half of the ELIC pore is occluded with Phe246 and Leu239 side-chains that narrow down the pore diameter to less than 1Å, while the remaining CP half is wide open (diameter of 6 Å). The two GLIC structures, on the other hand, are in the open state, being crystallized in the presence of an activating ligand proton. **Figure 5.5** compares the ELIC and GLIC structures after their optimal superimposition. In addition to a symmetric tilt of the pore forming helices, the most visible difference is a quaternary twist similar to that observed in nAChR. Bocquet et al. (Bocquet et al., 2009) reported that the lowest ENM mode, a quaternary

twist of the two domains, explains 29% of the structural difference between the cores of the structures. Overall these structural data are consistent with a model of pore opening involving a quaternary twist and tertiary deformation (Bocquet et al., 2009).

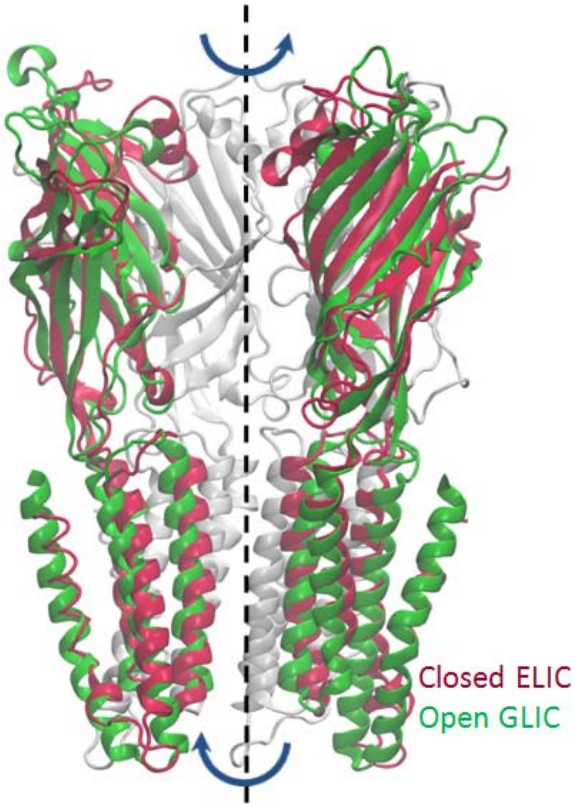


Figure 5.5 Comparison of open and closed bacterial ligand-gated ion channels (LGICs).

Comparison of bacterial homopentameric LGICs ELIC (2VL0) and GLIC (3EAM) shows the contribution of this quaternary twist mode to the conformational changes involved in activation. One subunit (closest to the viewer) is omitted to display the channel pore. Figure is adopted from (Bahar et al., 2009).

6.0 CONCLUSION AND FUTURE WORK

6.1 PROTEIN RECOGNITION DYNAMICS AND SMALL-MOLECULE BINDING

6.1.1 Why aren't some ANM modes observed in experimental datasets?

We presented the most comprehensive structural analyses of three drug target enzymes, HIV reverse transcriptase (RT), p38 MAP kinase, and cyclin-dependent kinase (Bakan & Bahar, 2009). Our results show that the conformational changes sampled by these enzymes upon binding a broad range of ligands closely correspond to those predicted by elastic network models. Despite the large size of these datasets, we should note that even when we consider the proteins at residue level, the degrees of freedom of the system (> 3,000 for RT, and > 1,000 for kinases) are much larger than the number of structures in the datasets. Hence, the PCA results that we showed are not expected to provide a complete picture of the conformational heterogeneity that these proteins display in solution.

We could say that we analyzed the therapeutically relevant part of the conformational space that is accessible to these targets. Biologically relevant structural changes (e.g., those accompanying binding of protein substrates, such as cyclin-bound Cdk2) were not included in our datasets. The inclusion of such structures could show a broader picture. Hence, it is likely

that some dominant solution modes are not observed in these X-ray datasets, due to their limited size and scope. It is also possible that missing modes are one of the top ranking ANM modes, that didn't find any counterpart among the PCA modes retrieved from experimental data.

p38 and Cdk2 share the same fold, so the slowest three ANM modes for these proteins are quite similar (compare **Figure 3.7** and **Figure 3.9**). It is notable that in the p38 ensemble, the ANM modes 1 and 3 were observed; whereas in the Cdk2-ensemble the ANM mode 2 was distinguished. The question that arises is: What may lead to such selective sampling of normal modes? Is it related to the differences in the sequence of these proteins, or is it related to the bound set of small molecules? It is one of our short term goals to find an answer to this question by expanding the dataset and carefully analyzing the sequence differences at hinge sites critical for slow motions. As mentioned above, we omitted certain protein kinases in our analyses e.g. cAMP dependent kinase which has over 80 structures in the PDB. A quick comparison of cAMP PCA modes with ANM modes showed ANM2 corresponds to PC1. We also omitted around 60 CyclinA2 bound Cdk2 structures, most of which also being inhibitor bound. PCA of this dataset showed that ANM1 is preferentially sampled by Cyclin A2 bound Cdk2. Comparing sequence features of cAMP, Cdk2, and p38, and maybe other kinases with more than 20 structures (there are over 15 different kinases with more than 20 structures), looks as a promising approach to understand why slow ANM modes are selectively sampled by this family of proteins. The sequence specificity was indeed not included in our computational models. As described above, ANM considers uniform masses for all residue types and uniform potentials for all pairwise interactions. Understanding this selective behavior may enable us to make educated choices

among slow ANM modes when generating alternative conformations rather than using slowest 10-20 ANM modes blindly.

In the HIV-1 RT case, the ANM calculations showed that the anti-correlated fluctuations of the fingers subdomain with respect to the RNase H domain (ANM1) is as likely to exist in solution as the anti-correlated fluctuations of fingers and thumb domain (ANM2 ~ PC1) (see **Figure 3.5**). Yet, we do not find among the extracted crystal structures, a mode closely corresponding to ANM1. This is presumably because NNRTI or DNA/RNA binding does not particularly stabilize a distinct configuration along this mode. The study of the solution dynamics of RT using NMR is beyond the limitations of this technique due the large size of the protein. Since we are only interested in the globular arrangement of RT subdomains, small angle x-ray scattering (SAXS) experiments seem useful. SAXS experiments provide low-resolution solution data about the shapes of proteins (Koch et al., 2003; Putnam et al., 2007). The shape information is in the form of a histogram of all pairwise distances in the protein. The method is particularly useful to identify structural changes of multi-domain proteins when such changes involve hinge motions (Forster et al., 2008). It is also a very simple task to generate distance distributions from protein structures deformed along ANM modes. We expect it to be a fruitful endeavor to seek for SAXS data on HIV-1 RT and compare them to our predictions.

6.1.2 Potential improvements to ANM-guided conformer generation

In MKP-3 studies, we utilized slow ANM modes to generate alternative MKP-3 conformations to be used in docking simulations. A combined ANM mode was introduced into an all-atom

energy minimization scheme in terms of α-carbon harmonic restraints. A few recent studies applied methods based on the same basic ideas to improve docking results. Cavasotto and Abagyan used all-atom normal modes to deform a cAMP-dependent kinase structure and performed energy minimization to restore the covalent geometry following deformation (Cavasotto et al., 2005). Floquet and Perahia introduced a single normal mode based restraint into energy minimization to generate conformations prior to docking (Floquet et al., 2006). May and Zacharias, on the other hand, relaxed the protein along the slowest 10 normal modes in their implementation of a coarse-grained docking protocol (May & Zacharias, 2008).

All of these distinct procedures and ours are prone to the limitations of downhill energy minimization. The normal mode restraints are typically introduced as weak forces compared to the covalent potentials and to the repulsive component of the LJ potential (Equation 2.4.4). During the energy minimization of a protein subject to a potential similar to Equation 2.5.2, interatomic contacts that are closer than certain thresholds (defined by atomic van der Waals radii) may result in strong repulsions, and therefore a very steep increase in the FF potential which may render NM restraints ineffective. In response to this strong interaction, the atoms may exhibit umphysical responses, i.e. the molecule starts to diverge in an improbable direction, or the side-chains can assume unrealistic rotameric configurations. We faced this problem when we used ANM-guided energy minimization to model a plausible structural transition for MKP-3 from its low-activity state to its high-activity state. These transitions are available in the MPEG video accessible at the publication web site of our article (http://www.nature.com/nchembio/journal/v5/n9/suppinfo/nchembio.190_S1.html). The first movie (Supplementary Movie 1) shows the fluctuations of MKP-3 along the combined ANM

mode. Snapshots for this movie were generated in around 40 iterative minimization steps. The second movie (Supplementary Movie 2) shows the complete transition of MKP-3 from low-activity to high-activity state. Over two hundred energy minimization steps were needed to reach the high-activity. As seen in the movie, the transition is rapid in the beginning and decelerates towards the end; and the Arg299 side-chain is distorted. The primary obstacle was the side-chain of Arg299. In its inactive state configuration, Arg299 side-chain clashed with the closing loop atoms and delayed the transition. In the high-activity state MKPs, Arg299 assumed a different rotameric state (compare two panels of **Figure 4.11**) which could enable an easier transition.

Such limitations can be overcome by using ANM restraints in a molecular dynamics simulation, whereby side-chains preventing the deformation or ANM-driven transition may jump to alternative rotameric states that would enable the movement along the ANM mode. This is in fact what Isin and Bahar developed in their study of rhodopsin (Isin et al., 2008). Another similar method, Normal Mode Following, was developed by Miloshevsky and Jordan (Miloshevsky & Jordan, 2006), in which the perturbations along normal modes are incorporated in a Monte Carlo simulation. The drawback of such Monte Carlo simulations is the duration of computations to reach distinct protein states. In molecular docking studies, one typically prefers to prepare the protein structure or the structural ensemble in less than an hour, and spend more time in performing the docking simulation. Spending a longer time, of the order of days, to sample protein conformations is not desirable feasible approach in most applications.

As a practical alternative, we may pursue NMA-guided energy minimization of the structure after temporary removal of all side-chains. Residues other than glycine and proline are

then mutated *in silico* to alanine, so as to facilitate smooth backbone transitions. After an ensemble of backbone conformations is prepared, side-chains can easily be modeled back into the newly generated structures using side-chain placement programs, such as SCWRL (Krivov et al., 2009). In certain cases, one may also achieve smooth transitions by only removing side-chains that make steric clashes. This may be especially convenient when using global/slow modes, in which most of the protein structure is preserved at its initial configuration. In MKP-3 case, for example, the conversion of the general acid loop and other binding site residues to alanine would enable an easy transition from the low-activity to the high-activity state. An automated procedure using Python programming and MMTK module of the PDB, as described above, may provide a powerful tool for flexible docking simulations.

6.2 MKP INHIBITORS

6.2.1 An iterative approach for designing more potent BCI analogs

We have developed models to aid in understanding the selectivity and the inhibitory mechanism of MKP inhibitors. Among those inhibitors that we studied, BCI is the most efficient and deserves further attention (ligand efficiency is defined as $\Delta g = [-RT \log(IC_{50})] / N_{heavy\text{-}atoms}$ (Hopkins et al., 2004); BCI (**Figure 4.6A)** is almost twice as efficient as PSI2106 (**Figure 4.3**). As we discussed in section 4.4, our model is yet not appropriate for assisting in structure-based analog design, due to the multiplicity of the likely solutions found at the putative binding site. This type of ligand design is usually based on a single binding mode of the ligand, often coming

from an X-ray co-crystal structure. In the absence of such a structure, our short-term goal is to identify the BCI binding mode that contributes most to the affinity of this compound.

Toward this aim, we would like to pursue an iterative design-synthesis strategy. Our collaborators in Dr. Day Laboratory have synthesized over 15 analogs of BCI. The inhibitory activity of these analogs has been assessed by Dr. Tsang Laboratory. These analogs bear peripheral additions or substitutions in the cyclohexylamino moiety. Our plan is to overlay these analogs onto the BCI docking poses and to establish the relationship between the activity of the compounds and their interactions with MKP-3.

As a supporting method, we also plan to adopt a more accurate method for binding free energy evaluation, so as to score the interactions of BCI and its analogs. As discussed above, typical docking scoring functions are simply additions of pairwise interaction terms (section 2.6). These terms represent the protein-ligand interactions *in vacuo*. Superior scoring methods that also account for the contribution of desolvation and electrostatic interactions in a more rigorous way are now becoming more conveniently available. One such method is the PB/SA scoring. The Poisson-Boltzmann (PB) part of this method (section 2.7.1) is an effective way to simulate the effects of water molecules in biological systems. The surface area (SA) part, on the other hand, accounts for desolvation of water from the hydrophobic surface at the interface between the protein and the ligand. This method is now implemented in DOCK 6 (Lang et al., 2009) and uses a fast electrostatics toolkit ZAP (Grant et al., 2001). It was also shown recently that, using ZAP libraries, the PB/SA scoring can be utilized in high-throughput calculations (Brown & Muchmore, 2006). We hypothesize that incorporating this more rigorous technique in an iterative

synthesis-design procedure will favorably aid in developing more potent BCI analogs for MKP inhibition.

6.2.2 Characterization of the putative BCI binding site

The shallow catalytic sites of MKPs make them a challenging target. The recent discovery of BCI and the putative BCI binding site opened new directions in targeting MKPs (Molina et al., 2009). This putative binding site also deserves more attention. We also aim at fully characterizing this site. ANM calculations showed that the general acid loop that makes up the putative site is flexible and has an intrinsic ability to close upon the catalytic cavity. We were also able to generate alternative loop configurations (**Figure 4.9**). Yet, what we do not know at this time are the weights of these alternative configurations. Although we have the crystal structure, this loop is affected by crystal contacts (**Figure 4.8**) and may not represent the most populated configuration in the absence of such contacts. Our plan is to further investigate the structure and dynamics of this loop using MD simulations. The establishment of the most populated (or lowest energy) configuration of this loop may help us setup virtual screening simulations for the identification of new compounds.

6.2.3 Estimating the druggability of the putative BCI binding site

BCI is a ~10 μM inhibitor of MKP3. Can we improve its potency? Certainly, but what is the maximal affinity achievable at the putative binding site? We would also like to answer this type

of questions. Recently, an MD simulation method for calculating druggability indices and identifying binding sites has been introduced (Seco et al., 2009). In this method, the protein is simulated in a mixture of water and isopropanol. Isopropanol is a small organic compound and diffuses fast in MD simulations. Hereby, it serves as a drug-like probe. Its distribution on the protein surface and cavities are converted into achievable affinities by a grid-based approach. The method may be advantageously employed toward the identification of novel binding pockets in addition to assessing the druggability of known pockets. We plan to apply this method to MKPs in an effort to estimate the maximal achievable affinity of the putative binding site. Further, we will seek other potential pockets that may remotely affect the flexibility of the general acid loop. If any, other novel pockets will also be tested by testing compounds identified in virtual screens.

6.2.4 Virtual screening for new chemotypes

After characterizing the BCI binding site, we plan to perform virtual high-throughput screening (vHTS) simulations for new chemotypes that work through the same allosteric mechanism as the BCI. We plan to incorporate the flexibility of the binding site in our screening. Pharmacophore based virtual screening methods are in general faster than docking. The dynamic pharmacophore model described by Carlson and McCammon seems particularly suitable for this task (Carlson et al., 2000). A dynamic model is generated using multiple target conformations from MD simulations. Pharmacophoric features and constraints are based on the conserved interactions between the probe molecules and the ensemble of conformations. In our case, we can use conserved solvent interactions from water-isopropanol mixture simulations of MKP-3. We will

use the Sybyl module Unity for 3D flexible pharmacophore based screening. We will also consider the ZINC database (Irwin & Shoichet, 2005) (http://zinc.docking.org/) of commercially available compounds for a more extensive screening. This database currently holds over 8 million drug-like compounds. The dynamic pharmacophore model will be used to select compounds that satisfy the desired features. The compounds derived from the initial search will be energy optimized and scored using PB/SA scoring function. The final set of hits from vHTS will be tested using zebrafish embryos by our collaborators in Dr. Tsang lab.

APPENDIX A

LIST OF PDB STRUCTURES USED IN ENSEMBLE ANALYSIS

Table A.1 PDB IDs of RT structures.

1BQM	1BQN	1C0T	1C0U	1C1B	1C1C	1DLO	1DTQ	1DTT	1EET
1EP4	1FK9	1FKO	1FKP	1HMV	1HNI	1HNV	1HPZ	1HQE	1HQU
1HYS	1IKV	1IKW	1IKX	1IKY	1J5O	1JKH	1JLA	1JLB	1JLC
1JLE	1JLF	1JLG	1JLQ	1KLM	1LW0	1LW2	1LWC	1LWE	1LWF
1N5Y	1N6Q	1QE1	1R0A	1REV	1RT1	1RT2	1RT3	1RT4	1RT5
1RT6	1RT7	1RTD	1RTH	1RTI	1RTJ	1S1T	1S1U	1S1V	1S1W
1S1X	1S6P	1S6Q	1S9E	1S9G	1SUQ	1SV5	1T03	1T05	1TKT
1TKX	1TKZ	1TL1	1TL3	1TV6	1TVR	1UWB	1VRT	1VRU	2B5J
2B6A	2BAN	2BE2	2HMI	2HND	2HNY	2HNZ	2I5J	2IAJ	2IC3
2OPP	2OPQ	2OPR	2OPS	2RF2	2RKI	2VG5	2VG6	2VG7	2ZD1
2ZE2	3BGR	3C6T	3C6U	3DI6	3DLE	3DLG	3DM2	3DMJ	3DOK
3DOL	3HVT								

Residues 4-63, 73-135, 142-443, 455-537, 1007-1064, 1068-1087, 1096-1212, 1233-1355, 1362-1427 were included in the analysis, corresponding to 89.4% of HIV-RT sequence. These structures were resolved at 3.5 Å resolution or higher. The reference structure is the unliganded form of RT determined by Esnouf et al. (Esnouf et al., 1995) (PDB id: 1RTJ). PDB IDs are colored according to scheme in Figure 3.1.

Table A.2 PDB IDs of p38 structures.

1A9U	1BL6	1BL7	1BMK	1DI9	1IAN	1KV1	1KV2	1LEW	1LEZ
1M7Q	1OUK	1OUY	1OVE	1OZ1	1P38	1R39	1R3C	1W7H	1W82
1W83	1W84	1WBN	1WBO	1WBS	1WBT	1WBV	1WBW	1WFC	1YQJ
1YW2	1YWR	1ZYJ	1ZZ2	1ZZL	2BAJ	2BAK	2BAL	2BAQ	2EWA
2FSL	2FSM	2FSO	2FST	2GFS	2GHL	2GHM	2GTM	2GTN	2I0H
2NPQ	2OKR	2OZA	2P5A	2PKJ	2PTJ	2PTO	2PUU	2PV5	2PV8
2QD9	2RG5	2RG6	2ZAZ	2ZB0	2ZB1	3BV2	3BV3	3BX5	3C5U
3CG2	3CTQ	3D7Z	3D83						

Residues 5-31, 36-114, 122-169, 185-351 were included in the analysis, corresponding to 89.2% of p38 MAP kinase sequence. These structures were resolved at 2.8 Å resolution or higher. The reference structure is the unliganded form of p38 determined by Wang et al. (Wang et al., 1997) (PDB id: 1P38). PDB IDs are colored according to scheme in Figure 3.1.

Table A.3 PDB IDs of Cdk2 structures.

1AQ1	1B38	1B39	1CKP	1DI8	1DM2	1E1V	1E1X	1FVT	1G5S
1GIH	1GII	1GIJ	1GZ8	1H00	1H01	1H07	1H08	1H0V	1H0W
1HCK	1HCL	1JSV	1JVP	1KE5	1KE6	1KE7	1KE8	1KE9	1OIQ
1OIR	1OIT	1P2A	1PW2	1PXI	1PXJ	1PXK	1PXL	1PXM	1PXN
1PXP	1PYE	1R78	1URW	1V1K	1VYZ	1W0X	1W8C	1WCC	1Y8Y
1Y91	1YKR	2A0C	2A4L	2B52	2B53	2B54	2B55	2BHE	2BHH
2BTR	2BTS	2C5Y	2C68	2C69	2C6I	2C6K	2C6L	2C6M	2C6O
2CLX	2DS1	2DUV	2EXM	2J9M	2R3F	2R3G	2R3H	2R3I	2R3J
2R3K	2R3L	2R3M	2R3N	2R3O	2R3P	2R3Q	2R3R	2R64	2UZN
2UZO	2V0D	2VTA	2VTH	2VTI	2VTJ	2VTL	2VTM	2VTN	2VTO
2VTP	2VTQ	2VTR	2VTS	2VV9	2W06				

Residues 1-35, 47-148, 164-296 were included in the analysis corresponding to 90.6% of Cdk2 sequence. These structures were resolved at 2.8 Å resolution or higher. The reference structure is the unliganded form of Cdk2 determined by Schulze-Gahmen et al. (Schulze-Gahmen et al., 1995) (PDB id: 1HCL). PDB IDs are colored according to scheme in Figure 3.1.

.

BIBLIOGRAPHY

Almo, S. C., Bonanno, J. B., Sauder, J. M., Emtage, S., Dilorenzo, T. P., Malashkevich, V. et al. (2007). Structural genomics of protein phosphatases. *J Struct.Funct.Genomics, 8,* 121-140.

Alonso, A., Sasin, J., Bottini, N., Friedberg, I., Friedberg, I., Osterman, A. et al. (2004). Protein tyrosine phosphatases in the human genome. *Cell, 117,* 699-711.

Altenbach, C., Cai, K., Klein-Seetharaman, J., Khorana, H. G., & Hubbell, W. L. (2001a). Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 65 in helix TM1 and residues in the sequence 306-319 at the cytoplasmic end of helix TM7 and in helix H8. *Biochemistry, 40,* 15483-15492.

Altenbach, C., Klein-Seetharaman, J., Cai, K., Khorana, H. G., & Hubbell, W. L. (2001b). Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 316 in helix 8 and residues in the sequence 60-75, covering the cytoplasmic end of helices TM1 and TM2 and their connection loop CL1. *Biochemistry, 40,* 15493-15500.

Altenbach, C., Kusnetzow, A. K., Ernst, O. P., Hofmann, K. P., & Hubbell, W. L. (2008). High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc.Natl.Acad.Sci.U.S.A, 105,* 7439-7444.

Altenbach, C., Yang, K., Farrens, D. L., Farahbakhsh, Z. T., Khorana, H. G., & Hubbell, W. L. (1996). Structural features and light-dependent changes in the cytoplasmic interhelical E-F loop region of rhodopsin: a site-directed spin-labeling study. *Biochemistry, 35,* 12470-12478.

Amadei, A., Ceruso, M. A., & Di, N. A. (1999). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins, 36,* 419-424.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J. et al. (1999). *LAPACK Users' Guide*. (3rd ed.) Philadelphia, PA: Society for Industrial and Applied Mathematics.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys.J, 80,* 505-515.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential. Folding & Design 2, 173-181. Ref Type: Journal (Full)

Bahar, I., Chennubhotla, C., & Tobi, D. (2007). Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr.Opin.Struct Biol, 17,* 633-640.

Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R., & Covell, D. G. (1999). Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J Mol.Biol, 285,* 1023-1037.

Bahar, I., Lezon, T. R., Bakan, A., & Shrivastava, I. H. (2009). Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chem.Rev.*.

Bahar, I. & Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Curr.Opin.Struct Biol, 15,* 586-592.

Bakan, A. & Bahar, I. (2009). The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc.Natl.Acad.Sci.U.S.A, 106,* 14349-14354.

Bakan, A., Lazo, J. S., Wipf, P., Brummond, K. M., & Bahar, I. (2008). Toward a molecular understanding of the interaction of dual specificity phosphatases with substrates: insights from structure-based modeling and high throughput screening. *Curr.Med.Chem., 15,* 2536-2544.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc.Natl.Acad.Sci.U.S.A, 98,* 10037-10041.

Barril, X. & Morley, S. D. (2005). Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J Med.Chem., 48,* 4432-4443.

Becker, O. M., Marantz, Y., Shacham, S., Inbal, B., Heifetz, A., Kalid, O. et al. (2004). G protein-coupled receptors: in silico drug discovery in 3D. *Proc.Natl.Acad.Sci.U.S.A, 101,* 11304-11309.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. et al. (2000). The Protein Data Bank. *Nucleic Acids Res., 28,* 235-242.

Best, R. B., Lindorff-Larsen, K., DePristo, M. A., & Vendruscolo, M. (2006). Relation between native ensembles and experimental structures of proteins. *Proc.Natl.Acad.Sci.U.S.A, 103,* 10901-10906.

Bhattacharya, S., Hall, S. E., & Vaidehi, N. (2008). Agonist-induced conformational changes in bovine rhodopsin: insight into activation of G-protein-coupled receptors. *J Mol.Biol., 382,* 539-555.

Bocquet, N., Nury, H., Baaden, M., Le, P. C., Changeux, J. P., Delarue, M. et al. (2009). X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. *Nature, 457,* 111-114.

Bottegoni, G., Kufareva, I., Totrov, M., & Abagyan, R. (2009). Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J Med.Chem., 52,* 397-406.

Bramson, H. N., Corona, J., Davis, S. T., Dickerson, S. H., Edelstein, M., Frye, S. V. et al. (2001). Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J.Med.Chem., 44,* 4339-4358.

Braun, W. & Go, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J Mol.Biol., 186,* 611-626.

Brejc, K., van Dijk, W. J., Klaassen, R. V., Schuurmans, M., van Der, O. J., Smit, A. B. et al. (2001). Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature, 411,* 269-276.

Brooijmans, N. & Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annu.Rev.Biophys.Biomol.Struct., 32,* 335-373.

Brooks, B. & Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc.Natl.Acad.Sci.U.S.A, 80,* 6571-6575.

Brooks, B. R., Brooks, C. L., III, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B. et al. (2009). CHARMM: the biomolecular simulation program. *J Comput Chem., 30,* 1545-1614.

Brown, S. P. & Muchmore, S. W. (2006). High-throughput calculation of protein-ligand binding affinities: modification and adaptation of the MM-PBSA protocol to enterprise grid computing. *J.Chem.Inf.Model., 46,* 999-1005.

Bush, B. L., Bayly, C. I., & Halgren, T. A. (1999). Consensus bond-charge increments fitted to electrostatic potential or field of many compounds: Application to MMFF94 training set. J Comput Chem. 20[14], 1495-1516.
Ref Type: Journal (Full)

Camps, M., Nichols, A., & Arkinstall, S. (2000). Dual specificity phosphatases: a gene family for control of MAP kinase function. *FASEB J, 14,* 6-16.

Camps, M., Nichols, A., Gillieron, C., Antonsson, B., Muda, M., Chabert, C. et al. (1998). Catalytic activation of the phosphatase MKP-3 by ERK2 mitogen-activated protein kinase. *Science, 280,* 1262-1265.

Carlson, H. A., Masukawa, K. M., Rubins, K., Bushman, F. D., Jorgensen, W. L., Lins, R. D. et al. (2000). Developing a dynamic pharmacophore model for HIV-1 integrase. *J Med.Chem., 43,* 2100-2114.

Cavasotto, C. N. & Abagyan, R. A. (2004). Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol.Biol, 337,* 209-225.

Cavasotto, C. N., Kovacs, J. A., & Abagyan, R. A. (2005). Representing receptor flexibility in ligand docking through relevant normal modes. *J Am.Chem.Soc., 127,* 9632-9640.

Cavasotto, C. N., Orry, A. J., & Abagyan, R. A. (2003). Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins, 51,* 423-433.

Cavasotto, C. N., Orry, A. J., Murgolo, N. J., Czarniecki, M. F., Kocsi, S. A., Hawes, B. E. et al. (2008). Discovery of novel chemotypes to a G-protein-coupled receptor through ligand-steered homology modeling and structure-based virtual screening. *J Med.Chem., 51,* 581-588.

Changeux, J. P. & Edelstein, S. J. (1998). Allosteric receptors after 30 years. *Neuron, 21,* 959-980.

Chen, H. Y., Yu, S. L., Chen, C. H., Chang, G. C., Chen, C. Y., Yuan, A. et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. *N.Engl.J Med., 356,* 11-20.

Chen, X., Ji, Z. L., & Chen, Y. Z. (2002). TTD: Therapeutic Target Database. *Nucleic Acids Res., 30,* 412-415.

Chen, X., Liu, M., & Gilson, M. K. (2001). BindingDB: a web-accessible molecular recognition database. *Comb.Chem.High Throughput.Screen., 4,* 719-725.

Cheng, X., Lu, B., Grant, B., Law, R. J., & McCammon, J. A. (2006). Channel opening motion of alpha7 nicotinic acetylcholine receptor as suggested by normal mode analysis. *J Mol.Biol., 355,* 310-324.

Chennubhotla, C. & Bahar, I. (2006). Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Mol.Syst.Biol., 2,* 36.

Chennubhotla, C., Yang, Z., & Bahar, I. (2008). Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol.Biosyst., 4,* 287-292.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S. et al. (2007). High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science, 318,* 1258-1265.

Congreve, M., Murray, C. W., & Blundell, T. L. (2005). Structural biology and drug discovery. *Drug Discov.Today, 10,* 895-907.

Corry, B. (2006). An energy-efficient gating mechanism in the acetylcholine receptor channel suggested by molecular and Brownian dynamics. *Biophys.J, 90,* 799-810.

Crozier, P. S., Stevens, M. J., Forrest, L. R., & Woolf, T. B. (2003). Molecular dynamics simulation of dark-adapted rhodopsin in an explicit membrane bilayer: coupling between local retinal and larger scale conformational change. *J Mol.Biol., 333,* 493-514.

Cui, Q. & Bahar, I. (2006). *Normal Mode Analysis: Theory and applications to biological and chemical systems*. (vols. 9) Boca Raton, FL: CRC Press.

Cymes, G. D. & Grosman, C. (2008). Pore-opening mechanism of the nicotinic acetylcholine receptor evinced by proton transfer. *Nat.Struct.Mol.Biol., 15,* 389-396.

Cymes, G. D., Ni, Y., & Grosman, C. (2005). Probing ion-channel pores one proton at a time. *Nature, 438,* 975-980.

Czajkowski, C. (2005). Neurobiology: triggers for channel opening. *Nature, 438,* 167-168.

Daylight (2007). Fingerprints - Screening and Similarity. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html [On-line].

Denu, J. M. & Dixon, J. E. (1998). Protein tyrosine phosphatases: mechanisms of catalysis and regulation. *Curr.Opin.Chem.Biol., 2,* 633-641.

Dobson, P. D. & Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat.Rev.Drug Discov., 7,* 205-220.

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res., 32,* W665-W667.

Doruker, P., Atilgan, A. R., & Bahar, I. (2000). Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins, 40,* 512-524.

Doruker, P., Jernigan, R. L., & Bahar, I. (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J Comput Chem., 23,* 119-127.

Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol.Biol., 230,* 543-574.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J.Comput.Aided Mol.Des, 11,* 425-445.

Esnouf, R., Ren, J., Ross, C., Jones, Y., Stammers, D., & Stuart, D. (1995). Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Nat.Struct Biol, 2,* 303-308.

Eyal, E., Yang, L. W., & Bahar, I. (2006). Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics., 22,* 2619-2627.

Farooq, A., Chaturvedi, G., Mujtaba, S., Plotnikova, O., Zeng, L., Dhalluin, C. et al. (2001). Solution structure of ERK2 binding domain of MAPK phosphatase MKP-3: structural insights into MKP-3 activation by ERK2. *Mol.Cell, 7,* 387-399.

Farooq, A., Plotnikova, O., Chaturvedi, G., Yan, S., Zeng, L., Zhang, Q. et al. (2003). Solution structure of the MAPK phosphatase PAC-1 catalytic domain. Insights into substrate-induced enzymatic activation of MKP. *Structure., 11,* 155-164.

Farooq, A. & Zhou, M. M. (2004). Structure and regulation of MAPK phosphatases. *Cell Signal., 16,* 769-779.

Farrens, D. L., Altenbach, C., Yang, K., Hubbell, W. L., & Khorana, H. G. (1996). Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science, 274,* 768-770.

Filmore, D. (2009). It's a GPCR world. Modern Drug Discovery 17, 24.
Ref Type: Journal (Full)

Floquet, N., Marechal, J. D., Badet-Denisot, M. A., Robert, C. H., Dauchez, M., & Perahia, D. (2006). Normal mode analysis as a prerequisite for drug design: application to matrix metalloproteinases inhibitors. *FEBS Lett., 580,* 5130-5136.

Forster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A., & Sali, A. (2008). Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. *J Mol.Biol., 382,* 1089-1106.

Fredriksson, R., Lagerstrom, M. C., Lundin, L. G., & Schioth, H. B. (2003). The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol.Pharmacol., 63,* 1256-1272.

Furukawa, T., Sunamura, M., Motoi, F., Matsuno, S., & Horii, A. (2003). Potential tumor suppressive pathway involving DUSP6/MKP-3 in pancreatic cancer. *Am.J Pathol., 162,* 1807-1815.

Ghanouni, P., Steenhuis, J. J., Farrens, D. L., & Kobilka, B. K. (2001). Agonist-induced conformational changes in the G-protein-coupling domain of the beta 2 adrenergic receptor. *Proc.Natl.Acad.Sci.U.S.A, 98,* 5997-6002.

Go, N., Noguti, T., & Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc.Natl.Acad.Sci.U.S.A, 80,* 3696-3700.

Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins, 8,* 195-202.

Grant, A. J., Pickup, B. T., & Nicholls, A. (2001). A smooth permittivity function for Poisson–Boltzmann solvation methods. J Comput Chem. 22, 608-640.
Ref Type: Journal (Full)

Grossfield, A., Feller, S. E., & Pitman, M. C. (2007). Convergence of molecular dynamics simulations of membrane proteins. *Proteins, 67,* 31-40.

Gsponer, J., Christodoulou, J., Cavalli, A., Bui, J. M., Richter, B., Dobson, C. M. et al. (2008). A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction. *Structure., 16,* 736-746.

Gunther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E. et al. (2008). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res., 36,* D919-D922.

Halgren, T. A. (1996a). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem. 17[5-6], 490-519.
Ref Type: Journal (Full)

Halgren, T. A. (1996b). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem. 17[5-6], 520-552.
Ref Type: Journal (Full)

Halgren, T. A. (1999a). MMFF VI. MMFF94s option for energy minimization studies. J Comput Chem. 20[7], 720-729.
Ref Type: Journal (Full)

Halgren, T. A. (1999b). MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. J Comput Chem. 20[7], 730-748.
Ref Type: Journal (Full)

Halgren, T. A. (1996c). Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. J Comput Chem. 17[5-6], 553-586.
Ref Type: Journal (Full)

Halgren, T. A. (1996d). Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem. 17[5-6], 616-641.
Ref Type: Journal (Full)

Halgren, T. A. & Nachbar, R. B. (1996). Merck molecular force field. IV. conformational energies and geometries for MMFF94. J Comput Chem. 17[5-6], 687-615.
Ref Type: Journal (Full)

Haliloglu, T., Bahar, I., & Erman, B. (1997). Gaussian dynamics of folded proteins. Phys 79, 3090-3093.
Ref Type: Journal (Full)

Hanson, M. A. & Stevens, R. C. (2009). Discovery of new GPCR biology: one receptor structure at a time. *Structure., 17,* 8-14.

Hilf, R. J. & Dutzler, R. (2008). X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature, 452,* 375-379.

Hilf, R. J. & Dutzler, R. (2009). Structure of a potentially open state of a proton-activated pentameric ligand-gated ion channel. *Nature, 457,* 115-118.

Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins, 33,* 417-429.

Hinsen, K. (2000). The molecular modeling toolkit: a new approach to molecular simulations. J.Comput.Chem. 21[2], 79-81.
Ref Type: Journal (Full)

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science, 268,* 1144-1149.

Hopkins, A. L., Groom, C. R., & Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discov.Today, 9,* 430-431.

Huang, S. Y. & Zou, X. (2007). Efficient molecular docking of NMR structures: application to HIV-1 protease. *Protein Sci., 16,* 43-51.

Hubbell, W. L., Altenbach, C., Hubbell, C. M., & Khorana, H. G. (2003). Rhodopsin structure, dynamics, and activation: a perspective from crystallography, site-directed spin labeling, sulfhydryl reactivity, and disulfide cross-linking. *Adv.Protein Chem., 63,* 243-290.

Hubbell, W. L., Cafiso, D. S., & Altenbach, C. (2000). Identifying conformational changes with site-directed spin labeling. *Nat.Struct.Biol., 7,* 735-739.

Huber, T., Botelho, A. V., Beyer, K., & Brown, M. F. (2004). Membrane model for the G-protein-coupled receptor rhodopsin: hydrophobic interface and dynamical structure. *Biophys.J, 86,* 2078-2100.

Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *J.Comput.Chem., 28,* 1145-1152.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *J Mol.Graph., 14,* 33-38.

Hung, A., Tai, K., & Sansom, M. S. (2005). Molecular dynamics simulation of the M2 helices within the nicotinic acetylcholine receptor transmembrane domain: structure and collective motions. *Biophys.J, 88,* 3321-3333.

Hutter, D., Chen, P., Barnes, J., & Liu, Y. (2000). Catalytic activation of mitogen-activated protein (MAP) kinase phosphatase-1 by binding to p38 MAP kinase: critical role of the p38 C-terminal domain in its negative regulation. *Biochem.J, 352,* 155-163.

Hyeon, C., Jennings, P. A., Adams, J. A., & Onuchic, J. N. (2009). Ligand-induced global transitions in the catalytic domain of protein kinase A. *Proc.Natl.Acad.Sci.U.S.A, 106,* 3023-3028.

Imming, P., Sinning, C., & Meyer, A. (2006). Drugs, their targets and the nature and number of drug targets. *Nat.Rev.Drug Discov., 5,* 821-834.

Irwin, J. J. & Shoichet, B. K. (2005). ZINC--a free database of commercially available compounds for virtual screening. *J Chem.Inf.Model., 45,* 177-182.

Isin, B., Rader, A. J., Dhiman, H. K., Klein-Seetharaman, J., & Bahar, I. (2006). Predisposition of the dark state of rhodopsin to functional changes in structure. *Proteins, 65,* 970-983.

Isin, B., Schulten, K., Tajkhorshid, E., & Bahar, I. (2008). Mechanism of signal propagation upon retinal isomerization: insights from molecular dynamics simulations of rhodopsin restrained by normal modes. *Biophys.J, 95,* 789-803.

Ivetac, A. & McCammon, J. A. (2009). Elucidating the Inhibition Mechanism of HIV-1 Non-Nucleoside Reverse Transcriptase Inhibitors through Multicopy Molecular Dynamics Simulations. *Journal of Molecular Biology, 388,* 644-658.

James, L. C., Roversi, P., & Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. *Science, 299,* 1362-1367.

James, L. C. & Tawfik, D. S. (2005). Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. *Proc.Natl.Acad.Sci.U.S.A, 102,* 12730-12735.

Jeffrey, K. L., Camps, M., Rommel, C., & Mackay, C. R. (2007). Targeting dual-specificity phosphatases: manipulating MAP kinase signalling and immune responses. *Nat.Rev.Drug Discov., 6,* 391-403.

Jeong, D. G., Cho, Y. H., Yoon, T. S., Kim, J. H., Ryu, S. E., & Kim, S. J. (2007). Crystal structure of the catalytic domain of human DUSP5, a dual specificity MAP kinase protein phosphatase. *Proteins, 66,* 253-258.

Jeong, D. G., Yoon, T. S., Kim, J. H., Shim, M. Y., Jung, S. K., Son, J. H. et al. (2006). Crystal structure of the catalytic domain of human MAP kinase phosphatase 5: structural insight into constitutively active phosphatase. *J Mol.Biol, 360,* 946-955.

Johnston, P. A., Foster, C. A., Shun, T. Y., Skoko, J. J., Shinde, S., Wipf, P. et al. (2007). Development and implementation of a 384-well homogeneous fluorescence intensity high-throughput screening assay to identify mitogen-activated protein kinase phosphatase-1 dual-specificity protein phosphatase inhibitors. *Assay.Drug Dev.Technol., 5,* 319-332.

Jolliffe, I. T. (2002). *Principal Component Analysis*. (2nd ed.) (vols. XXIX) New York: Springer.

Jones, G., Willett, P., & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol.Biol., 245,* 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol.Biol, 267,* 727-748.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science, 303,* 1813-1818.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A, 32,* 922-923.

Kenakin, T. (2003). Ligand-selective receptor conformations revisited: the promise and the problem. *Trends Pharmacol.Sci., 24,* 346-354.

Kensch, O., Restle, T., Wohrl, B. M., Goody, R. S., & Steinhoff, H. J. (2000). Temperature-dependent equilibrium between the open and closed conformation of the p66 subunit of HIV-1 reverse transcriptase revealed by site-directed spin labelling. *J Mol.Biol, 301,* 1029-1039.

Kern, D. & Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology, 13,* 748-757.

Kitao, A. & Go, N. (1999). Investigating protein dynamics in collective coordinate space. *Curr.Opin.Struct Biol, 9,* 164-169.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat.Rev.Drug Discov., 3,* 935-949.

Kobilka, B. K. (2007). G protein coupled receptor structure and activation. *Biochim.Biophys.Acta, 1768,* 794-807.

Koch, M. H., Vachette, P., & Svergun, D. I. (2003). Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q.Rev.Biophys., 36,* 147-227.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., & Steitz, T. A. (1992). Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science, 256,* 1783-1790.

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc.Natl.Acad.Sci.U.S.A, 44,* 98-104.

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins, 77,* 778-795.

Kumar, S., Boehm, J., & Lee, J. C. (2003). p38 MAP kinases: key signalling molecules as therapeutic targets for inflammatory diseases. *Nat.Rev.Drug Discov., 2,* 717-726.

Kusnetzow, A. K., Altenbach, C., & Hubbell, W. L. (2006). Conformational states and dynamics of rhodopsin in micelles and bilayers. *Biochemistry, 45,* 5538-5550.

Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V. et al. (2009). DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA., 15,* 1219-1230.

Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S. et al. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science, 320,* 1471-1475.

Lazo, J. S., Nunes, R., Skoko, J. J., Queiroz de Oliveira, P. E., Vogt, A., & Wipf, P. (2006). Novel benzofuran inhibitors of human mitogen-activated protein kinase phosphatase-1. *Bioorg.Med.Chem., 14,* 5643-5650.

Lazo, J. S., Skoko, J. J., Werner, S., Mitasev, B., Bakan, A., Koizumi, F. et al. (2007). Structurally unique inhibitors of human mitogen-activated protein kinase phosphatase-1 identified in a pyrrole carboxamide library. *J Pharmacol.Exp.Ther., 322,* 940-947.

Leach, A. R. (2001). *Molecular Modeling: Principles and Applications*. (2nd ed.) Essex, England: Prentice Hall.

Lemaitre, V., Yeagle, P., & Watts, A. (2005). Molecular dynamics simulations of retinal in rhodopsin: from the dark-adapted state towards lumirhodopsin. *Biochemistry, 44,* 12667-12680.

Levin, E. J., Kondrashov, D. A., Wesenberg, G. E., & Phillips, G. N., Jr. (2007). Ensemble refinement of protein crystal structures: validation and application. *Structure., 15,* 1040-1052.

Levitt, M., Sander, C., & Stern, P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol.Biol., 181,* 423-447.

Liu, S., Sun, J. P., Zhou, B., & Zhang, Z. Y. (2006). Structural basis of docking interactions between ERK2 and MAP kinase phosphatase 3. *Proc.Natl.Acad.Sci.U.S.A, 103,* 5326-5331.

Liu, X., Xu, Y., Li, H., Wang, X., Jiang, H., & Barrantes, F. J. (2008). Mechanics of channel gating of the nicotinic acetylcholine receptor. *PLoS.Comput Biol., 4,* e19.

Lovell, S. C., Word, J. M., Richardson, J. S., & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins, 40,* 389-408.

Lu, M. & Ma, J. (2005). The role of shape in determining molecular motions. *Biophys.J, 89,* 2395-2401.

Ma, B., Kumar, S., Tsai, C. J., & Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng, 12,* 713-720.

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure., 13,* 373-380.

Marques, O. & Sanejouand, Y. H. (1995). Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins, 23,* 557-560.

May, A. & Zacharias, M. (2008). Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J Med.Chem., 51,* 3499-3506.

Meng, E. C., Shoichet, B. K., & Kuntz, I. D. (1992). Automated docking with grid-based energy evaluation. J.Comput.Chem. 13, 505-524.
Ref Type: Journal (Full)

Miloshevsky, G. V. & Jordan, P. C. (2006). The open state gating mechanism of gramicidin a requires relative opposed monomer rotation and simultaneous lateral displacement. *Structure., 14,* 1241-1249.

Mittermaier, A. & Kay, L. E. (2006). New Tools Provide New Insights in NMR Studies of Protein Dynamics. *Science, 312,* 224-228.

Miyazawa, A., Fujiyoshi, Y., & Unwin, N. (2003). Structure and gating mechanism of the acetylcholine receptor pore. *Nature, 423,* 949-955.

Molina, G., Vogt, A., Bakan, A., Dai, W., Queiroz de, O. P., Znosko, W. et al. (2009). Zebrafish chemical screening reveals an inhibitor of Dusp6 that expands cardiac cell lineages. *Nat.Chem.Biol., 5,* 680-687.

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J.Mol.Biol., 12,* 88-118.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. et al. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem. 19[14], 1639-1662.
Ref Type: Journal (Full)

Morris, G. M., Goodsell, D. S., Huey, R., & Olson, A. J. (1996). Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J.Comput.Aided Mol.Des, 10,* 293-304.

Naik, V. M., Krimm, S., Denton, J. B., Nemethy, G., & Scheraga, H. A. (1984). Vibrational analysis of peptides, polypeptides and proteins. XXVII. Structure of gramicidin S from normal mode analyses of low-energy conformations. *Int.J Pept.Protein Res., 24,* 613-626.

Nicolay, S. & Sanejouand, Y. H. (2006). Functional modes of proteins are among the most robust. *Phys.Rev.Lett., 96,* 078104.

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008). Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem.Cent.J, 2,* 5.

Okazaki, K. & Takada, S. (2008). Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc.Natl.Acad.Sci.U.S.A, 105,* 11182-11187.

Oldham, W. M. & Hamm, H. E. (2008). Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat.Rev.Mol.Cell Biol., 9,* 60-71.

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nat.Rev.Drug Discov., 5,* 993-996.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A. et al. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science, 289,* 739-745.

Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W., & Ernst, O. P. (2008). Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature, 454,* 183-187.

Pavletich, N. P. (1999). Mechanisms of cyclin-dependent kinase regulation: structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors. *J Mol.Biol, 287,* 821-828.

Peleg, G., Ghanouni, P., Kobilka, B. K., & Zare, R. N. (2001). Single-molecule spectroscopy of the beta(2) adrenergic receptor: observation of conformational substates in a membrane protein. *Proc.Natl.Acad.Sci.U.S.A, 98,* 8469-8474.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. et al. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem., 25,* 1605-1612.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E. et al. (2005). Scalable molecular dynamics with NAMD. *J Comput Chem., 26,* 1781-1802.

Prasad, J. C., Goldstone, J. V., Camacho, C. J., Vajda, S., & Stegeman, J. J. (2007). Ensemble modeling of substrate binding to cytochromes P450: analysis of catalytic differences between CYP1A orthologs. *Biochemistry, 46,* 2640-2654.

Puius, Y. A., Zhao, Y., Sullivan, M., Lawrence, D. S., Almo, S. C., & Zhang, Z. Y. (1997). Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. *Proc.Natl.Acad.Sci.U.S.A, 94,* 13420-13425.

Putnam, C. D., Hammel, M., Hura, G. L., & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q.Rev.Biophys., 40,* 191-285.

Rader, A. J., Anderson, G., Isin, B., Khorana, H. G., Bahar, I., & Klein-Seetharaman, J. (2004). Identification of core amino acids stabilizing rhodopsin. *Proc.Natl.Acad.Sci.U.S.A, 101,* 7246-7251.

Ragno, R., Frasca, S., Manetti, F., Brizzi, A., & Massa, S. (2005). HIV-reverse transcriptase inhibition: inclusion of ligand-induced fit by cross-docking studies. *J Med.Chem., 48,* 200-212.

Resek, J. F., Farahbakhsh, Z. T., Hubbell, W. L., & Khorana, H. G. (1993). Formation of the meta II photointermediate is accompanied by conformational changes in the cytoplasmic surface of rhodopsin. *Biochemistry, 32,* 12025-12032.

Reynolds, K. A., Katritch, V., & Abagyan, R. (2009). Identifying conformational changes of the beta(2) adrenoceptor that enable accurate prediction of ligand/receptor interactions and screening for GPCR modulators. *J Comput Aided Mol.Des, 23,* 273-288.

Rohrig, U. F., Guidoni, L., & Rothlisberger, U. (2002). Early steps of the intramolecular signal transduction in rhodopsin explored by molecular dynamics simulations. *Biochemistry, 41,* 10799-10809.

Ruvinsky, A. M. (2007). Calculations of protein-ligand binding entropy of relative and overall molecular motions. *J Comput Aided Mol.Des, 21,* 361-370.

Saam, J., Tajkhorshid, E., Hayashi, S., & Schulten, K. (2002). Molecular dynamics investigation of primary photoinduced events in the activation of rhodopsin. *Biophys.J, 83,* 3097-3112.

Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol.Biol., 234,* 779-815.

Scheerer, P., Park, J. H., Hildebrand, P. W., Kim, Y. J., Krauss, N., Choe, H. W. et al. (2008). Crystal structure of opsin in its G-protein-interacting conformation. *Nature, 455,* 497-502.

Schulze-Gahmen, U., Brandsen, J., Jones, H. D., Morgan, D. O., Meijer, L., Vesely, J. et al. (1995). Multiple modes of ligand recognition: crystal structures of cyclin-dependent protein kinase 2 in complex with ATP and two inhibitors, olomoucine and isopentenyladenine. *Proteins, 22,* 378-391.

Seco, J., Luque, F. J., & Barril, X. (2009). Binding site detection and druggability index from first principles. *J Med.Chem., 52,* 2363-2371.

Sen, T. Z. & Jernigan, R. L. (2006). Optimizing the parameters of the Gaussian Network Model for ATP binding proteins. In Q.Cui & I. Bahar (Eds.), (Boca Raton, FL: Chapman & Hall/CRC Mathematical & Computational Biology.

Shacham, S., Marantz, Y., Bar-Haim, S., Kalid, O., Warshaviak, D., Avisar, N. et al. (2004). PREDICT modeling and in-silico screening for G-protein coupled receptors. *Proteins, 57,* 51-86.

Shoichet, B. K., McGovern, S. L., Wei, B., & Irwin, J. J. (2002). Lead discovery using molecular docking. *Curr.Opin.Chem.Biol., 6,* 439-446.

Showalter, S. A. & Bruschweiler, R. (2007). Quantitative Molecular Ensemble Interpretation of NMR Dipolar Couplings without Restraints. *Journal of the American Chemical Society, 129,* 4158-4159.

Sine, S. M. & Engel, A. G. (2006). Recent advances in Cys-loop receptor structure and function. *Nature, 440,* 448-455.

Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A., & Sansom, M. S. (1996). HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J Mol.Graph., 14,* 354-60, 376.

Song, G. & Jernigan, R. L. (2006). An enhanced elastic network model to represent the motions of domain-swapped proteins. *Proteins, 63,* 197-209.

Stahlberg, H., Fotiadis, D., Scheuring, S., Remigy, H., Braun, T., Mitsuoka, K. et al. (2001). Two-dimensional crystals: a powerful approach to assess structure, function and dynamics of membrane proteins. *FEBS Lett., 504,* 166-172.

Stewart, A. E., Dowd, S., Keyse, S. M., & McDonald, N. Q. (1999). Crystal structure of the MAPK phosphatase Pyst1 catalytic domain and implications for regulated activation. *Nat.Struct Biol, 6,* 174-181.

Sullivan, S. M. & Holyoak, T. (2008). Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proc.Natl.Acad.Sci.U.S.A, 105,* 13829-13834.

Swaminath, G., Xiang, Y., Lee, T. W., Steenhuis, J., Parnot, C., & Kobilka, B. K. (2004). Sequential binding of agonists to the beta2 adrenoceptor. Kinetic evidence for intermediate conformational states. *J Biol.Chem., 279,* 686-691.

Szarecka, A., Xu, Y., & Tang, P. (2007). Dynamics of heteropentameric nicotinic acetylcholine receptor: implications of the gating mechanism. *Proteins, 68,* 948-960.

Taly, A., Delarue, M., Grutter, T., Nilges, M., Le, N. N., Corringer, P. J. et al. (2005). Normal mode analysis suggests a quaternary twist model for the nicotinic receptor gating mechanism. *Biophys.J, 88,* 3954-3965.

Tama, F. & Brooks, C. L. (2006). Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu.Rev.Biophys.Biomol.Struct, 35,* 115-133.

Tama, F. & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng, 14,* 1-6.

Tang, C., Schwieters, C. D., & Clore, G. M. (2007). Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature, 449,* 1078-1082.

Tao, X. & Tong, L. (2007). Crystal structure of the MAP kinase binding domain and the catalytic domain of human MKP5. *Protein Sci., 16,* 880-886.

Temiz, N. A. & Bahar, I. (2002). Inhibitor binding alters the directions of domain motions in HIV-1 reverse transcriptase. *Proteins, 49,* 61-70.

Theodosiou, A. & Ashworth, A. (2002). MAP kinase phosphatases. *Genome Biol, 3,* REVIEWS3009.

Thisse, B. & Thisse, C. (2005). Functions and regulations of fibroblast growth factor signaling during embryonic development. *Dev.Biol., 287,* 390-402.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R. et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res., 13,* 2129-2141.

Tobi, D. & Bahar, I. (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc.Natl.Acad.Sci.U.S.A, 102,* 18908-18913.

Tonks, N. K. (2006). Protein tyrosine phosphatases: from genes, to function, to disease. *Nat.Rev.Mol.Cell Biol., 7,* 833-846.

Totrov, M. & Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr.Opin.Struct Biol, 18,* 178-184.

Turjanski, A. G., Gutkind, J. S., Best, R. B., & Hummer, G. (2008). Binding-Induced Folding of a Natively Unstructured Transcription Factor. *PLoS Comput Biol, 4,* e1000060.

Unwin, N. (2005). Refined structure of the nicotinic acetylcholine receptor at 4A resolution. *J Mol.Biol., 346,* 967-989.

Unwin, N. (1995). Acetylcholine receptor channel imaged in the open state. *Nature, 373,* 37-43.

Varady, J., Wu, X., Fang, X., Min, J., Hu, Z., Levant, B. et al. (2003). Molecular modeling of the three-dimensional structure of dopamine 3 (D3) subtype receptor: discovery of novel and potent D3 ligands through a hybrid pharmacophore- and structure-based database searching approach. *J Med.Chem., 46,* 4377-4392.

Vicent, S., Garayoa, M., Lopez-Picazo, J. M., Lozano, M. D., Toledo, G., Thunnissen, F. B. et al. (2004). Mitogen-activated protein kinase phosphatase-1 is overexpressed in non-small cell lung cancer and is an independent predictor of outcome in patients. *Clin.Cancer Res., 10,* 3639-3649.

Vogt, A., Cooley, K. A., Brisson, M., Tarpley, M. G., Wipf, P., & Lazo, J. S. (2003). Cell-active dual specificity phosphatase inhibitors identified by high-content screening. *Chem.Biol., 10,* 733-742.

Vogt, A., McDonald, P. R., Tamewitz, A., Sikorski, R. P., Wipf, P., Skoko, J. J., III et al. (2008). A cell-active inhibitor of mitogen-activated protein kinase phosphatases restores paclitaxel-induced apoptosis in dexamethasone-protected cancer cells. *Mol.Cancer Ther., 7,* 330-340.

Vogt, A., Tamewitz, A., Skoko, J., Sikorski, R. P., Giuliano, K. A., & Lazo, J. S. (2005). The benzo[c]phenanthridine alkaloid, sanguinarine, is a selective, cell-active inhibitor of mitogen-activated protein kinase phosphatase-1. *J Biol Chem., 280,* 19078-19086.

Wang, H. Y., Cheng, Z., & Malbon, C. C. (2003). Overexpression of mitogen-activated protein kinase phosphatases MKP1, MKP2 in human breast cancer. *Cancer Lett., 191,* 229-237.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *J Comput Chem., 25,* 1157-1174.

Wang, Z., Harkins, P. C., Ulevitch, R. J., Han, J., Cobb, M. H., & Goldsmith, E. J. (1997). The structure of mitogen-activated protein kinase p38 at 2.1-A resolution. *Proc.Natl.Acad.Sci.U.S.A, 94,* 2327-2332.

Werner, S., Iyer, P. S., Fodor, M. D., Coleman, C. M., Twining, L. A., Mitasev, B. et al. (2006). Solution-phase synthesis of a tricyclic pyrrole-2-carboxamide discovery library applying a stetter-Paal-Knorr reaction sequence. *J Comb.Chem., 8,* 368-380.

White, A., Pargellis, C. A., Studts, J. M., Werneburg, B. G., & Farmer, B. T. (2007). Molecular basis of MAPK-activated protein kinase 2:p38 assembly. *Proc.Natl.Acad.Sci.U.S.A, 104,* 6353-6358.

Wiesmann, C., Barr, K. J., Kung, J., Zhu, J., Erlanson, D. A., Shen, W. et al. (2004). Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat.Struct Mol.Biol, 11,* 730-737.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D. et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res., 36,* D901-D906.

Wu, G. S. (2007). Role of mitogen-activated protein kinase phosphatases (MKPs) in cancer. *Cancer Metastasis Rev., 26,* 579-585.

Yang, L., Song, G., Carriquiry, A., & Jernigan, R. L. (2008). Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure., 16,* 321-330.

Yang, L., Song, G., & Jernigan, R. L. (2007). How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys.J., 93,* 920-929.

Yang, L., Tan, C. H., Hsieh, M. J., Wang, J., Duan, Y., Cieplak, P. et al. (2006). New-generation amber united-atom force field. *J Phys.Chem.B, 110,* 13166-13176.

Yang, L. W., Eyal, E., Bahar, I., & Kitao, A. (2009). Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. *Bioinformatics., 25,* 606-614.

Zhang, Q., Muller, M., Chen, C. H., Zeng, L., Farooq, A., & Zhou, M. M. (2005). New insights into the catalytic activation of the MAPK phosphatase PAC-1 induced by its substrate MAPK ERK2 binding. *J Mol.Biol, 354,* 777-788.

Zheng, C. J., Han, L. Y., Yap, C. W., Ji, Z. L., Cao, Z. W., & Chen, Y. Z. (2006a). Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol.Rev., 58,* 259-279.

Zheng, W., Brooks, B. R., & Thirumalai, D. (2006b). Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc.Natl.Acad.Sci.U.S.A, 103,* 7664-7669.

Zhou, B., Wu, L., Shen, K., Zhang, J., Lawrence, D. S., & Zhang, Z. Y. (2001). Multiple regions of MAP kinase phosphatase 3 are involved in its recognition and activation by ERK2. *J Biol Chem., 276,* 6506-6515.

Zhou, B., Zhang, J., Liu, S., Reddy, S., Wang, F., & Zhang, Z. Y. (2006). Mapping ERK2-MKP3 binding interfaces by hydrogen/deuterium exchange mass spectrometry. *J Biol Chem., 281,* 38834-38844.

Zhou, Z., Madrid, M., Evanseck, J. D., & Madura, J. D. (2005). Effect of a bound non-nucleoside RT inhibitor on the dynamics of wild-type and mutant HIV-1 reverse transcriptase. *J Am.Chem.Soc., 127,* 17253-17260.